EXTENDED HOMOZYGOSITY SCORE TESTS TO DETECT POSITIVE SELECTION IN GENOME-WIDE SCANS

A Dissertation

by

MING ZHONG

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

May 2010

Major Subject: Statistics

EXTENDED HOMOZYGOSITY SCORE TESTS TO DETECT POSITIVE SELECTION IN GENOME-WIDE SCANS

A Dissertation

by

MING ZHONG

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Approved by:

Chair of Committee,	Ruzong Fan
Committee Members,	Michael Longnecker
	Raymond J. Carroll
	Clare Gill
	John Charles Huber Jr.
Head of Department,	Simon J. Sheather

May 2010

Major Subject: Statistics

ABSTRACT

Extended Homozygosity Score Tests to Detect
Positive Selection in Genome-wide Scans. (May 2010)
Ming Zhong, B.S., University of Science & Technology of China;
M.S., Texas A&M University
Chair of Advisory Committee: Ruzong Fan

Positive natural selection is recognized as the driving force underneath evolution. One of the surest signatures of recent positive selection is a local elevation of advantageous allele frequency and linkage disequilibrium (LD). This dissertation proposes a new test statistic to detect excess homozygosity based on a simple counting measure, which serves as a surrogate indicator of recent positive selection. Three tests are developed upon the new measure: (a) an extended genotype-based homozygosity test (EGHT), (b) a hidden Markov model test (HMMT), and (c) an extended haplotype-based homozygosity test (EHHT). The null hypotheses of all three tests assume random mating and Hardy-Weinberg equilibrium (HWE). They differ in how to treat LD under H_0. The EGHT assumes linkage equilibrium (LE) besides HWE while the EHHT allows arbitrary multi-locus LD. The HMMT stands between these two extremes and assumes pairwise but no higher-order disequilibrium interactions. We first conduct simulation study to compare the three score tests and verify that the EHHT is the most conservative one. We compare the performance of the EHHT with the prevailing detection methods and the EHHT has higher or similar power. We also evaluate the impact of simple demographic history on the EHHT and the simulation study suggests that the EHHT is resistant to the false-positive confounders resulting from simple demographic models. After extensive simulation studies, all three tests are then applied on HapMap Phase II data and we are able to replicate find-

ings reported in the literature. We can also identify new candidate regions that may undergo recent selection through a set of filtering criteria including highest EHHT scores, high derived allele frequency and large population differentiation. Finally, we propose a cross-population comparison test statistic to detect chromosome regions in which there is no significant excess homozygosity in one population but homozygosity remains high in another population.

To Yuhong, Jingneng and my parents

ACKNOWLEDGMENTS

First, I am deeply indebted to my advisor, Dr. Ruzong Fan, for his guidance, patience and support in the Ph.D. program. Ruzong has been a friend and mentor. He was always there to listen and to share his own experience, knowledge, and give advice. He taught me different ways to approach a research problem, and beyond that, the persistence required to accomplish any goal. I will benefit from the working experience with him for the rest of my life.

My special thanks to Dr. Michael Longnecker for his consistent support over the years. Besides serving on my committee, Michael is a role model in terms of teaching and working.

My thanks are extended to Dr. Raymond J. Carroll, Dr. Clare Gill, and Dr. John Charles Huber Jr., for serving on my committee and providing me with valuable comments. I also want to thank Dr. Fred Dahm for recruiting me to this department in 2005. I really cherish this opportunity to pursue my interests.

I would like to thank Marilyn Randall, Sandra Wood and Shou-Fen Lee for making the department a very welcoming place.

Finally, and most importantly, I am deeply grateful to my lovely wife Yuhong. Her tolerance of my occasional bad moods and unwavering love helped me to face the most difficult time and start my graduate career on the right foot. I can't thank her enough for all the encouragement, trust, hope and love she brought me.

TABLE OF CONTENTS

LIST OF TABLES

TABLE Page

TABLE Page

TABLE Page

LIST OF FIGURES

FIGURE Page

CHAPTER I

INTRODUCTION

In population genetics, positive selection can lead to an increase in the frequency of an advantageous allele. The selected allele may arise rapidly enough that there's no time for recombination to eliminate its association with nearby polymorphisms. Neutral and nearly neutral genetic variation linked to it will also become more prevalent, which is often called genetic hitchhiking effect. Recent positive selection may introduce a selective sweep that will reduce or eliminate genetic variation. Consequently, a reduction in total genetic variation results in less opportunity for recombination and lead to high levels of LD in the vicinity of the trait gene (Bamshad and Wooding, 2003; Hudson et al., 1994; Kim and Nielsen, 2004; Ronald and Akey, 2005; Vallender and Lahn, 2004).

A lot of detection methods have been proposed to locate candidate regions and a few widely adopted approaches are actually based on the same measure: Extended Haplotype Homozygosity (EHH) (Sabeti et al., 2002; Voight et al., 2006; Hanchard et al., 2006; Sabeti et al., 2007). EHH is defined as the probability that two randomly chosen chromosomes are homozygous at all SNPs between A and B, inclusively. Explicitly, if the N chromosomes in a sample form G homozygous groups, with each group i having n_i elements, EHH is defined as

$$EHH = \frac{\sum_{i=1}^{G} \binom{n_i}{2}}{\binom{N}{2}} \tag{1.1}$$

From the definition, we can see that low haplotype diversity leads to an EHH value

---

The journal model is *Behavior Genetics*.

close to 1. Although this probability measure of homozygosity is proved to be insightful, the methods developed upon it suffer from the fact that the related underlying true distribution is not clear. In order to carry out these tests, the truncation and/or log-transformation steps need to be applied in the data processing stage.

An extended stretch of homozygosity can serve as a surrogate indicator of recent positive selection. Here we present three new homozygosity score tests to detect positive selection in genome-wide scans. In Chapter II, we propose a simple measure of homozygosity based on counting the consecutive homozygotes around the core single nucleotide polymorphism (SNP) and construct a test statistic to evaluate excess homozygosity. From there we develop three tests including: (a) an extended genotype-based homozygosity test (EGHT), (b) a hidden Markov model test (HMMT), and (c) an extended haplotype-based homozygosity test (EHHT). The null hypothesis for the EGHT postulates both HWE and LE whereas the EHHT explicitly takes into account multi-locus LD. The HMMT occupies the intermediate ground of allowing for pairwise LD. We derive the detailed statistical algorithms to implement these tests.

In Chapter III, we first investigate their false positive rates conditioning on the existing allele frequencies, i.e., under the null hypothesis of the EGHT. The comparison verifies that the EHHT is the most conservative among three tests. We then study the authentic type I error rates of the EHHT by running coalescent simulation to generate SNP samples allowing multi-locus interaction. As the most conservative and robust of all three, the EHHT is compared with current methods in terms of power. We also run coalescent simulations to evaluate the impact of population demography on the EHHT. We apply these tests to the widely studied HapMap Phase II data. Our results are consistent with previous findings across the genome and within specific candidate regions. In addition, we propose to identify new candidate regions through a set of filtering criteria. In the end, consider the situations that there is

no significant excess homozygosity in one population but homozygosity remains high in another population, we propose a cross-population comparison score test. The summary and conclusions are presented in Chapter IV.

CHAPTER II

METHOD

Consider a random sample of n unrelated individuals typed on a large number of SNPs. Upon defining the core SNP as SNP 0, we can denote the SNPs around the core SNP 0 as $k = \cdots, -2, -1, 0, 1, 2, \cdots$. Here one may need to do truncations if the core SNP 0 is close to or on the boundary. Let M be the indicator of whether the SNP 0 is homozygous, let L, R be the number of consecutive homozygous SNPs flanking the core SNP from the left, right side, respectively. If the core SNP 0 is heterozygous ($M = 0$), by convention we define $L = R = 0$. The extent of homozygosity is then measured by the total $T = L + M + R$. The quantities L, M, R, and T are random variables varying from individual to individual. If we can get the estimate of the mean μ and variance σ^2 of T, then we can conduct a test for excess homozygosity. More precisely, let T_j be the value of T for person j in the random sample. Based on the central limit theorem, the score statistic

$$S = \frac{1}{\sqrt{n\sigma^2}} \sum_{j=1}^{n} (T_j - \mu) \tag{2.1}$$

should be approximately standard normal. Because we are concerned with excess homozygosity, a one-sided test is appropriate.

We then introduce three new tests: an extended genotype-based homozygosity test, a hidden Markov model test and an extended haplotype-based homozygosity test. Before we derive the mean and the variance of each test statistic, let us consider their corresponding null hypotheses. In each instance, the null hypothesis includes random mating and hence global HWE. Thus, two phased-haplotypes H_1/H_2 with frequencies h_1 and h_2 are transmitted to the offsprings with frequency $2h_1h_2$ when $H_1 \neq H_2$ and

with frequency h_1^2 when $H_1 = H_2$. Only the null hypothesis of EGHT invokes the further assumption of LE where h_1 and h_2 equal the product of the underlying allele frequencies. Under the null hypothesis of the HMMT, the SNPs exhibit pairwise but not higher-order LD. For the EHHT, arbitrary LD is allowed. In summary, the null hypotheses of the three tests are

- Null hypothesis of EGHT: HWE and LE;
- Null hypothesis of HMMT: HWE and pairwise LD but no higher-order disequilibrium interactions;
- Null hypothesis of EHHT: HWE and arbitrary multi-locus LD.

In human genome, LD tends to extend the stretch of homozygosity surrounding a central marker given high density SNPs such as in the HapMap Phase II data. The mean μ calculated for the EGHT is too small since LE is assumed under the null hypothesis. Consequently, there are too many false positives favoring selection. As the other extreme, the EHHT tends to condition on existing haplotype diversity and is very conservative. The HMMT stands between these extremes and conditions on pairwise LD. Given the ubiquity of pairwise disequilibrium, this seems like a reasonable compromise.

Regardless of the test, one can decompose the theoretical mean of T as $\mu = \mathrm{E}\,(L) + \mathrm{E}\,(M) + \mathrm{E}\,(R)$. Because M is an indicator random variable, $\mathrm{E}\,(M) = \Pr(M = 1)$ and $\mathrm{Var}(M) = \Pr(M = 1)[1 - \Pr(M = 1)]$. If we let X_k be the unordered genotype of SNP k, then it is natural to calculate $\mathrm{E}\,(L)$ as $\mathrm{E}\,[\mathrm{E}\,(L \mid X_0)]$. Because X_0 takes only three possible values, the outer expectation in $\mathrm{E}\,[\mathrm{E}\,(L \mid X_0)]$ is trivial to compute. The case $X_0 = 1/2$ is easiest of all because $L = 0$ when $X_0 = 1/2$ and $M = 0$. Similar comments apply to $\mathrm{E}\,(R)$. The most natural route to calculate the variance σ^2 follows

the formula

$$\begin{aligned}\mathrm{Var}(T) &= \mathrm{Var}(L) + \mathrm{Var}(M) + \mathrm{Var}(R) \\ &\quad +2\,\mathrm{Cov}(L,M) + 2\,\mathrm{Cov}(L,R) + 2\,\mathrm{Cov}(M,R) \qquad (2.2)\end{aligned}$$

Again it is productive to condition on X_0. For instance,

$$\begin{aligned}\mathrm{Var}(L) &= \mathrm{Var}[\mathrm{E}\,(L \mid X_0)] + \mathrm{E}\,[\mathrm{Var}(L \mid X_0)], \qquad (2.3)\\ \mathrm{Var}(R) &= \mathrm{Var}[\mathrm{E}\,(R \mid X_0)] + \mathrm{E}\,[\mathrm{Var}(R \mid X_0)], \qquad (2.4)\end{aligned}$$

and, assuming L and R are independent given X_0,

$$\begin{aligned}\mathrm{Cov}(L,R) &= \mathrm{Cov}[\mathrm{E}\,(L \mid X_0), \mathrm{E}\,(R \mid X_0)] + \mathrm{E}\,[\mathrm{Cov}(L,R \mid X_0)] \\ &= \mathrm{Cov}[\mathrm{E}\,(L \mid X_0), \mathrm{E}\,(R \mid X_0)]. \qquad (2.5)\end{aligned}$$

It is also worth pointing out that $\mathrm{E}\,(LM) = \mathrm{E}(L)$ and $\mathrm{E}\,(RM) = \mathrm{E}(R)$, since L and R equal 0 when M does, and when $M = 1$, LM equals L and RM equals R. Thus, one has

$$\begin{aligned}\mathrm{Cov}(L,M) &= \mathrm{E}\,(LM) - \mathrm{E}\,(L)\,\mathrm{E}\,(M) = \mathrm{E}\,(L)[1 - \mathrm{E}\,(M)], \qquad (2.6)\\ \mathrm{Cov}(R,M) &= \mathrm{E}\,(RM) - \mathrm{E}\,(R)\,\mathrm{E}\,(M) = \mathrm{E}\,(R)[1 - \mathrm{E}\,(M)]. \qquad (2.7)\end{aligned}$$

These considerations emphasize the importance of finding the distributions of L and R conditional on $X_0 = 1/1$ and $X_0 = 2/2$. The next few sections tackle this issue.

A. The distribution of L and R under the null hypothesis of EGHT

Under the dual assumptions of HWE and LE, the conditional distributions of the random variables L and R depend only on M instead of the particular value of X_0. Let p_{k1} and p_{k2} be the frequencies of the two alleles at SNP k. In this notation one

can readily deduce that

$$\begin{aligned} \Pr(M=1) &= p_{01}^2 + p_{02}^2, && (2.8) \\ \Pr\Big(R \geq r \mid M=1\Big) &= \prod_{k=1}^{r}(p_{k1}^2 + p_{k2}^2), && (2.9) \\ \Pr\Big(L \geq \ell \mid M=1\Big) &= \prod_{k=-\ell}^{-1}(p_{k1}^2 + p_{k2}^2), && (2.10) \end{aligned}$$

where the products are empty when $r = 0$ or $\ell = 0$. In practice, one can either estimate the allele frequencies p_{k1} and p_{k2} from the sample or substitute known values for them. To compute the conditional means and variances of L and R numerically, we recommend the right-tail sums

$$\mathrm{E}\,(Y) = \sum_{j=1}^{\infty}\Pr(Y \geq j), \qquad \mathrm{E}\,(Y^2) = \sum_{j=1}^{\infty}(2j-1)\Pr(Y \geq j), \qquad (2.11)$$

valid for any nonnegative random variable Y with integer values. The sums defining $\mathrm{E}\,(Y)$ and $\mathrm{E}\,(Y^2)$ can be truncated as soon as they stabilize.

B. The distribution of L and R under the null hypothesis of HMMT

To find the conditional distributions of L and R under this scenario, we run a Markov chain whose states are the three unordered SNP genotypes 1/1, 1/2, and 2/2 and whose epochs are SNPs. If we again suppose that SNP 0 is the central SNP, then the genotype sequence $\ldots, X_{-1}, X_0, X_1, \ldots$ constitutes the chain. Every SNP emits a signal, either a 1 for a homozygote or a 0 for a heterozygote. Assuming pairwise LD but no higher-order linkage interactions, the two sections of the chain to the left and right of the central SNP are independent conditional on the state X_0 at that SNP. The only nontrivial states X_0 that come into play at SNP 0 are 1/1 and 2/2, and these occur with the Hardy-Weinberg probabilities p_{01}^2 and p_{02}^2, respectively.

To compute the conditional mean and variance of R, it suffices to compute the probabilities $\Pr(R \geq r \mid X_0)$. This can be achieved by running Baum's forward algorithm for an infinite sequence of emitted 1's. One pass of the algorithm is adequate. When SNP r is visited, $\Pr(R \geq r \mid X_0)$ becomes available. This description omits mention of transition probabilities. Along either haplotype, the transition from allele j at SNP r to allele k at SNP $r+1$ is governed by the known conditional probabilities that explicitly account for pairwise LD. These conditional probabilities can be readily estimated from sample data. We traverse the left and right sections in opposite directions, so their transition probabilities must take this into account.

Under the assumption of no genotyping error, the complexities of the hidden Markov chain can be replaced by simple recurrence relations. Let $p_{r,j\to k}$ be the LD probability that allele j at locus r is followed by allele k at locus $r+1$ on a chromosome segment containing both loci. If we also let

$$a_{rj} \;=\; \Pr(X_r = j/j, R \geq r \mid X_0 = 1/1), \tag{2.12}$$

then we can write $\Pr(R \geq r \mid X_0 = 1/1) = a_{r1} + a_{r2}$. Thus, computing the conditional mean and variance of R reduces to the problem of computing the a_{r1} and a_{r2}. By convention we take $a_{01} = 1$ and $a_{02} = 0$. These choices lead to the recurrences

$$\begin{aligned}
a_{r1} &= \Pr(X_r = 1/1, R \geq r \mid X_0 = 1/1) \\
&= \Pr(X_r = 1/1, R \geq r-1, X_{r-1} = 1/1 \text{ or } 2/2 \mid X_0 = 1/1) \\
&= \Pr(R \geq r-1, X_{r-1} = 1/1 \mid X_0 = 1/1) \Pr(X_r = 1/1 \mid X_{r-1} = 1/1) \\
&\quad + \Pr(R \geq r-1, X_{r-1} = 2/2 \mid X_0 = 1/1) \Pr(X_r = 1/1 \mid X_{r-1} = 2/2) \\
&= a_{r-1,1}\,(p_{r-1,1\to 1})^2 + a_{r-1,2}\,(p_{r-1,2\to 1})^2, \\
a_{r2} &= a_{r-1,1}\,(p_{r-1,1\to 2})^2 + a_{r-1,2}\,(p_{r-1,2\to 2})^2.
\end{aligned} \tag{2.13}$$

Computation of the vector $a_r = (a_{r1}, a_{r2})$ should continue until

$$(2r-1)\Pr(R \geq r \mid X_0 = 1/1) \;<\; \epsilon \tag{2.14}$$

for $\epsilon > 0$ suitably small. In order to compute $\Pr(R \geq r \mid X_0 = 2/2)$, we similarly define

$$b_{rj} \;=\; \Pr(R \geq r, X_r = j/j \mid X_0 = 2/2). \tag{2.15}$$

The b_{rj} satisfy exactly the same recurrences as the a_{rj} but differ in the initial values ($b_{01} = 0$ and $b_{02} = 1$).

As just stated, the distribution of L is conditionally independent of R given X_0. We can develop similar recurrent relationship for $\Pr(L \geq \ell \mid X_0 = 1/1)$ and $\Pr(L \geq \ell \mid X_0 = 2/2)$. Let $c_{\ell j}$ be the probability that $X_{-\ell} = j/j$ and $L \geq \ell$ given $X_0 = 1/1$. The conventions $c_{01} = 1$ and $c_{02} = 0$ are consistent with the formula $\Pr(L \geq \ell \mid X_0 = 1/1) = c_{\ell 1} + c_{\ell 2}$. Furthermore, we have the recurrences

$$\begin{aligned}
c_{\ell 1} &= \Pr(X_{-\ell} = 1/1, L \geq \ell \mid X_0 = 1/1) \\
&= \Pr(X_{-\ell} = 1/1, L \geq \ell - 1, X_{-\ell+1} = 1/1 \text{ or } 2/2 \mid X_0 = 1/1) \\
&= \Pr(L \geq \ell - 1, X_{-\ell+1} = 1/1 \mid X_0 = 1/1) \\
&\quad \times \Pr(X_{-\ell} = 1/1 \mid L \geq \ell - 1, X_{-\ell+1} = 1/1, X_0 = 1/1) \\
&\quad + \Pr(L \geq \ell - 1, X_{-\ell+1} = 2/2 \mid X_0 = 1/1) \\
&\quad \times \Pr(X_{-\ell} = 1/1 \mid L \geq \ell - 1, X_{-\ell+1} = 2/2, X_0 = 1/1) \\
&= c_{\ell-1,1} \Pr(X_{-\ell} = 1/1 \mid X_{-\ell+1} = 1/1) + c_{\ell-1,2} \Pr(X_{-\ell} = 1/1 \mid X_{-\ell+1} = 2/2) \\
&= \frac{\Pr(X_{-\ell} = 1/1)}{\Pr(X_{-\ell+1} = 1/1)} c_{\ell-1,1} \left(p_{-\ell,1\to 1}\right)^2 + \frac{\Pr(X_{-\ell} = 1/1)}{\Pr(X_{-\ell+1} = 2/2)} c_{\ell-1,2} \left(p_{-\ell,1\to 2}\right)^2, \\
c_{\ell 2} &= \frac{\Pr(X_{-\ell} = 2/2)}{\Pr(X_{-\ell+1} = 1/1)} c_{\ell-1,1} \left(p_{-\ell,2\to 1}\right)^2 + \frac{\Pr(X_{-\ell} = 2/2)}{\Pr(X_{-\ell+1} = 2/2)} c_{\ell-1,2} \left(p_{-\ell,2\to 2}\right)^2
\end{aligned} \tag{2.16}$$

If we let $d_{\ell j}$ be the probability that $L \geq \ell$ and $X_{-\ell} = j/j$ given $X_0 = 2/2$, the same recurrences as the $c_{\ell j}$ also hold for $d_{\ell j}$, but the initial values are given by $d_{01} = 0$ and $d_{02} = 1$.

C. The distribution of L and R under the null hypothesis of EHHT

In the presence of arbitrary LD, the fast recurrences (2.13) and (2.16) for a_{rj} and b_{rj} no longer apply. However, if we define $h_{i_0,\ldots,i_r}$ to be the population frequency of the haplotype $(i_0, \ldots, i_r)$ extending from SNP 0 to SNP r, then the formula

$$\Pr(R \geq r \mid X_0 = i_0/i_0) \;=\; \frac{1}{p^2_{0,i_0}} \sum_{i_1=1}^{2} \cdots \sum_{i_r=2}^{2} h^2_{i_0,\ldots,i_r} \tag{2.17}$$

delivers the required right-tail probabilities. When all conceivable haplotypes are possible, there are 2^r terms in the multiple sum, and the formula as it stands is cumbersome. On the other hand, if only a few haplotypes are possible, then the sum is straightforward to evaluate. The moment formulas (2.11) are still applicable. The haplotype frequencies $h_{i_0,\ldots,i_r}$ can be estimated from genotype data by the EM algorithm (Ayers and Lange, 2008; Long et al., 1995).

D. Cross-population EHHT test

In this section, we propose a cross-population comparison test statistic, XP-EHHT, to detect chromosome regions in which there is no significant excess homozygosity in one population but homozygosity remains high in another population.

The XP-EHHT is defined with respect to two populations, A and B, at a given core SNP. First, only the SNPs for which there are data for both populations A and B are selected as cores SNPs. For population A, let us denote the mean and variance of T by μ_A and σ^2_A, respectively; similarly, let μ_B and σ^2_B be the mean and variance

of T for the population B, respectively.

Next, we restrict our attention to the chromosome region around the core SNP to calculate its EHHT score in population A. Consider a random sample of n_A unrelated individuals of population A typed on a large number of SNPs around the core SNP 0. Let T_{Ai} be the value of T for person i in the random sample of population A. The summation $\sum_{i=1}^{n_A} T_{Ai}$ provides a measurement of total homozygosity in the sample of population A. If there is no significant excess homozygosity in the sample of population A, $\sum_{i=1}^{n_A}[T_{Ai} - \mu_A]$ tends to be close to 0; otherwise, it tends to be much larger than 0, which serves as an indication of excess homozygosity around the core SNP 0. We can proceed analogously with respect to another population B. Hence, in order to test the excess homozygosity of one population against the other, the pooled-test statistic is defined as

$$S_{AB} = \frac{\sum_{i=1}^{n_A}(T_{Ai} - \mu_A)/n_A - \sum_{j=1}^{n_B}(T_{Bj} - \mu_B)/n_B}{\sigma_P^2\sqrt{1/n_A + 1/n_B}}, \tag{2.18}$$

where $\sigma_P^2 = \sqrt{\frac{(n_A-1)\sigma_A^2+(n_B-1)\sigma_B^2}{n_A+n_B-2}}$ is the pooled-variance of random variable T for populations A and B. A pooled-test score S_{AB} is directional: a positive score suggests excess homozygosity in population A, whereas a negative score suggests excess homozygosity in population B. In practice, the mean and variance parameters μ_A, μ_B, σ_A^2 and σ_B^2 needs to be estimated by empirical data. By doing so, the test statistic S_{AB} would follow a t-distribution with a degree of freedom of $n_A + n_B - 2$. For Hapmap II data, the sample sizes are big enough,$n_A + n_B$ is equal or larger than 120, S_{AB} should approximately follow a standard normal distribution according to central limit theorem.

CHAPTER III

RESULTS

A. Type I error rates

By construction, the EHHT is the most conservative one among the three tests of EGHT, HMMT and EHHT. As a confirmation, we perform type I error comparison by simulating genotype data under the null of EGHT, i.e., assuming HWE and LE.

We generate genotype data according to the allele frequencies from the CHB+JPT and YRI samples on Chromosome 1, 2 and 15. Table I summarizes the performance of the EGHT statistic over 10^8 random samples of $n = 60$, 90, 125, 250, 500, and 1250 individuals. The results in Table I suggest that false positive rates are appropriate for $\alpha = 0.05$ and 0.01 when $n \geq 60$. For $\alpha = 0.001$, false positive rates are too high when $n \leq 250$ and close to the nominal level when $n \geq 500$. Finally for $\alpha = 0.0001$, it takes a sample size of at least $n = 1250$ to get a false positive rate close to the nominal level.

Table II reports false positive rates for the HMMT with data simulated under the same conditions. We also add results for the CEU sample. Under the unrealistic null hypothesis appropriate to the EGHT, the HMMT is too conservative. False positive rates are smaller than the nominal level for $\alpha = 0.05$, 0.01, 0.001, 0.0001, and 0.00001 when sample size is 60 or greater. For $\alpha = 0.000001$, false positive rates are reasonable when $n \geq 125$.

Table I. Type I error rates of the extended genotype-based homozygosity test (EGHT). All results based on 10^8 simulations and the HapMap Phase II SNP allele frequencies.

Sample Size n	**Chr.**	**Population**	**Type I Error Rates When**			
			$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.001$	$\alpha = 0.0001$
60	1	CHB+JPT	0.04938718	0.01035739	0.00135778	0.00022274
		YRI	0.04937440	0.01028255	0.00132343	0.00021185
	2	CHB+JPT	0.04940133	0.01034034	0.00135832	0.00022680
		YRI	0.04947162	0.01030993	0.00132751	0.00021625
	15	CHB+JPT	0.04932188	0.01031996	0.00135838	0.00022544
		YRI	0.04940568	0.01031936	0.00132825	0.00021288
90	1	CHB+JPT	0.04952430	0.01020326	0.00123662	0.00018285
		YRI	0.04955716	0.01019437	0.00122318	0.00017691
	2	CHB+JPT	0.04955975	0.01022223	0.00124360	0.00018541
		YRI	0.04964012	0.01020800	0.00123110	0.00017911
	15	CHB+JPT	0.04960285	0.01022962	0.00124236	0.00018548
		YRI	0.04964285	0.01022434	0.00122597	0.00017911
125	1	CHB+JPT	0.04968785	0.01014673	0.00117489	0.00016028
		YRI	0.04974835	0.01015954	0.00116450	0.00015561
	2	CHB+JPT	0.04973875	0.01018076	0.00118329	0.00016391
		YRI	0.04964882	0.01011992	0.00116113	0.00015746
	15	CHB+JPT	0.04966775	0.01012681	0.00116167	0.00015570
		YRI	0.04972462	0.01014865	0.00115829	0.00015481
250	1	CHB+JPT	0.04983445	0.01006635	0.00108067	0.00012857
		YRI	0.04988641	0.01009465	0.00108552	0.00012828
	2	CHB+JPT	0.04989234	0.01009633	0.00109923	0.00013297
		YRI	0.04989319	0.01007496	0.00108337	0.00012691
	15	CHB+JPT	0.04979736	0.01008825	0.00109559	0.00013371
		YRI	0.04982615	0.01007379	0.00108525	0.00012952
500	1	CHB+JPT	0.04988070	0.01003214	0.00104432	0.00011518
		YRI	0.04992955	0.01003731	0.00104643	0.00011390
	2	CHB+JPT	0.04988924	0.01003926	0.00103772	0.00011406
		YRI	0.04992835	0.01004388	0.00103968	0.00011400
	15	CHB+JPT	0.04995869	0.01001675	0.00103760	0.00011289
		YRI	0.04992122	0.01004642	0.00104140	0.00011416
1250	1	CHB+JPT	0.04992160	0.01000125	0.00101399	0.00010860
		YRI	0.05001056	0.01002057	0.00101685	0.00010386
	2	CHB+JPT	0.04990222	0.00998176	0.00102016	0.00010494
		YRI	0.04992216	0.01001738	0.00101428	0.00010371
	15	CHB+JPT	0.04996471	0.01004044	0.00102895	0.00010903
		YRI	0.04991000	0.01000108	0.00101769	0.00010529

Table II. Type I error rates of the hidden Markov model tests (HMMT). All results based on 10^8 simulations and the HapMap Phase II SNP allele frequencies.

Sample Size n	**Chr.**	**Population**	**Type I Error Rates When** $\alpha =$					
			0.05	0.01	0.001	0.0001	0.00001	0.000001
60	1	CEU	0.0326	0.0040	0.00024	2.7e-05	5.5e-06	1.4e-06
		CHB+JPT	0.0337	0.0043	0.00029	3.9e-05	9.0e-06	2.8e-06
		YRI	0.0318	0.0038	0.00021	2.0e-05	3.0e-06	5.9e-07
	2	CEU	0.0324	0.0040	0.00025	2.9e-05	6.1e-06	1.9e-06
		CHB+JPT	0.0332	0.0043	0.00031	5.0e-05	1.5e-05	6.2e-06
		YRI	0.0317	0.0037	0.00020	2.0e-05	2.8e-06	4.9e-07
	15	CEU	0.0325	0.0040	0.00026	3.4e-05	7.9e-06	2.7e-06
		CHB+JPT	0.0331	0.0042	0.00028	4.0e-05	1.2e-05	6.3e-06
		YRI	0.0317	0.0037	0.00020	1.8e-05	2.6e-06	5.5e-07
90	1	CEU	0.0313	0.0041	0.00024	2.2e-05	2.9e-06	4.1e-07
		CHB+JPT	0.0327	0.0043	0.00027	2.7e-05	4.7e-06	1.3e-06
		YRI	0.0305	0.0039	0.00022	1.9e-05	2.4e-06	4.0e-07
	2	CEU	0.0315	0.0041	0.00025	2.3e-05	2.9e-06	4.5e-07
		CHB+JPT	0.0320	0.0043	0.00028	3.0e-05	5.3e-06	1.3e-06
		YRI	0.0305	0.0039	0.00022	1.8e-05	2.5e-06	4.1e-07
	15	CEU	0.0315	0.0042	0.00026	2.8e-05	4.8e-06	1.3e-06
		CHB+JPT	0.0320	0.0043	0.00027	2.9e-05	6.2e-06	2.4e-06
		YRI	0.0306	0.0039	0.00022	1.8e-05	2.3e-06	4.3e-07
125	1	CEU	0.0308	0.0042	0.00026	2.2e-05	2.6e-06	4.0e-07
		CHB+JPT	0.0315	0.0044	0.00028	2.4e-05	3.1e-06	5.6e-07
		YRI	0.0299	0.0039	0.00023	1.7e-05	1.9e-06	3.4e-07
	2	CEU	0.0310	0.0042	0.00026	2.2e-05	2.4e-06	3.5e-07
		CHB+JPT	0.0315	0.0044	0.00029	2.7e-05	3.3e-06	6.4e-07
		YRI	0.0299	0.0039	0.00023	1.8e-05	1.7e-06	2.1e-07
	15	CEU	0.0310	0.0043	0.00027	2.4e-05	2.7e-06	4.3e-07
		CHB+JPT	0.0315	0.0044	0.00028	2.7e-05	4.2e-06	1.2e-06
		YRI	0.0301	0.0040	0.00024	1.9e-05	2.2e-06	2.6e-07
250	1	CEU	0.0301	0.0043	0.00029	2.3e-05	2.2e-06	2.4e-07
		CHB+JPT	0.0307	0.0044	0.00030	2.5e-05	2.5e-06	3.4e-07
		YRI	0.0292	0.0041	0.00026	1.9e-05	1.6e-06	2.0e-07
	2	CEU	0.0305	0.0044	0.00030	2.3e-05	2.5e-06	2.6e-07
		CHB+JPT	0.0307	0.0045	0.00031	2.6e-05	2.7e-06	3.4e-07
		YRI	0.0293	0.0041	0.00026	2.1e-05	1.9e-06	2.5e-07
	15	CEU	0.0305	0.0044	0.00029	2.3e-05	2.2e-06	3.1e-07
		CHB+JPT	0.0309	0.0045	0.00031	2.5e-05	2.5e-06	2.5e-07
		YRI	0.0294	0.0041	0.00027	2.1e-05	2.3e-06	3.2e-07

Table III reports false positive rates for the EHHT under the null hypothesis of the EGHT. Because of the high computational demand of EHHT, 10^5 replicates instead of 10^8 are performed to calculate a type I error rate. False positive rates are smaller than the nominal level for $\alpha = 0.05$, 0.01, and 0.001 when sample size is 60 or greater. At the nominal level $\alpha = 0.0001$, the false positive rates are reasonable when sample size is 60 or greater. Comparing with the results in Table I, the false

positive rates for the EHHT are lower than those of the EGHT. At the nominal level of 0.05, the false positive rates for the EHHT are also lower than those of the HMMT in Table II.

Table III. Type I error rates of the extended haplotype-based homozygosity test (EHHT). All results based on 10^5 simulations and the HapMap Phase II SNP allele frequencies.

Sample Size n	**Chr.**	**Population**	**Type I Error Rates When** $\alpha =$			
			0.05	0.01	0.001	0.0001
60	1	CEU	0.02271	0.00373	0.00047	0.00009
		CHB+JPT	0.02407	0.00441	0.00056	0.00012
		YRI	0.02331	0.00382	0.00053	0.00010
	2	CEU	0.02586	0.00447	0.00037	0.00010
		CHB+JPT	0.02561	0.00477	0.00078	0.00011
		YRI	0.02364	0.00344	0.00039	0.00006
	15	CEU	0.02349	0.00386	0.00041	0.00003
		CHB+JPT	0.02445	0.00404	0.00045	0.00008
		YRI	0.02308	0.00373	0.00034	0.00006
90	1	CEU	0.02326	0.00353	0.00032	0.00006
		CHB+JPT	0.02371	0.00389	0.00032	0.00001
		YRI	0.02209	0.00373	0.00046	0.00005
	2	CEU	0.02446	0.00387	0.00032	0.00006
		CHB+JPT	0.02433	0.00463	0.00060	0.00009
		YRI	0.02454	0.00366	0.00023	0.00003
	15	CEU	0.02336	0.00361	0.00049	0.00006
		CHB+JPT	0.02424	0.00367	0.00038	0.00008
		YRI	0.02324	0.00343	0.00030	0.00007
125	1	CEU	0.02323	0.00336	0.00023	0.00002
		CHB+JPT	0.02433	0.00394	0.00037	0.00005
		YRI	0.02208	0.00324	0.00027	0.00001
	2	CEU	0.02264	0.00355	0.00036	0.00006
		CHB+JPT	0.02379	0.00423	0.00054	0.00004
		YRI	0.02316	0.00296	0.00017	0.00004
	15	CEU	0.02362	0.00364	0.00032	0.00002
		CHB+JPT	0.02393	0.00416	0.00060	0.00008
		YRI	0.02345	0.00335	0.00028	0.00002
250	1	CEU	0.02259	0.00350	0.00017	0.00002
		CHB+JPT	0.02447	0.00392	0.00034	0.00004
		YRI	0.02149	0.00318	0.00017	0.00002
	2	CEU	0.02439	0.00334	0.00027	0.00002
		CHB+JPT	0.02380	0.00397	0.00026	0.00002
		YRI	0.02336	0.00348	0.00020	0.00001
	15	CEU	0.02325	0.00352	0.00022	0.00000
		CHB+JPT	0.02424	0.00379	0.00037	0.00006
		YRI	0.02200	0.00328	0.00029	0.00002

In conclusion, the EHHT is the most conservative one among the three tests. Hereafter, we focus on the performance evaluation of EHHT.

B. Coalescent simulations on type I error rates of EHHT

In a coalescent simulation, the random genealogy of the sample is first generated and then mutations are randomly placed on the genealogy.

We first use SelSim to generate SNP sequences under the neutral model of Spencer and Coop (2004). A genomic region containing m SNPs is simulated 5,000 times to obtain the type I error rates and m ranges from 51 to 101 in order to evaluate the impact of SNP sequence length on the EHHT performance. In addition, uniform recombination rates of $\rho = 1.5, 3, 6$ and 9 between SNPs are considered. The type I error rates at two nominal levels $\alpha = 0.05$ and $\alpha = 0.01$ are reported in Table IV, which are the proportion of the EHHT scores of the central SNP that exceeds 95th and 99th percentiles of the standard normal. We notice that when the number of SNPs was 51, the type I error is inflated and much bigger than the nominal levels, for any of the four recombination rates. However, the type I error rates drop fast as the number of SNP increases. Once the number of SNPs reaches 71, the type I error rates stabilize. It indicates that during the homozygote counting procedure, the truncation at the boundary due to short SNP sequence introduces bias and causes a problem of high false positives. Fortunately, almost all contemporary genomic data comprise thousands of, or even millions of, SNPs. Based on the results of Table IV, the type I error rates are lower than or around the nominal level except for the recombination rate $\rho = 1.5$ when the number of SNPs is larger or equal to the 71. When $\rho = 1.5$, the type I error rates are slightly higher than the nominal levels. Therefore, the simulation demonstrates that the EHHT is conservative and it has appropriate type I error rates when applied to SNPs which are reasonably far away from the boundary (≥ 35).

Table IV. Type I error rates of the extended haplotype-based homozygosity test (EHHT). All results are based on 5,000 simulations using software SelSim.

Sample Size n	**# of SNPs**	ρ	**Nominal Level** α=0.05	**Nominal Level** α=0.01	**# of SNPs**	ρ	**Nominal Level** α=0.05	**Nominal Level** α=0.01
100	51	1.5	0.1496	0.0592	61	1.5	0.0832	0.0260
		3	0.1268	0.0486		3	0.0548	0.0164
		6	0.1066	0.0388		6	0.0462	0.0130
		9	0.1036	0.0362		9	0.0464	0.0124
	71	1.5	0.0592	0.0182	81	1.5	0.0562	0.0148
		3	0.0392	0.0092		3	0.0374	0.0080
		6	0.0362	0.0062		6	0.0272	0.0052
		9	0.0296	0.0042		9	0.0306	0.0048
	91	1.5	0.0566	0.0184	101	1.5	0.0578	0.0200
		3	0.0414	0.0102		3	0.0412	0.0104
		6	0.0260	0.0052		6	0.0326	0.0056
		9	0.0262	0.0054		9	0.0274	0.0052
60	91	1.5	0.0580	0.0208	101	1.5	0.0626	0.0226
		3	0.0344	0.0084		3	0.0396	0.0112
		6	0.0298	0.0052		6	0.0330	0.0084
		9	0.0252	0.0062		9	0.0248	0.0052

C. Power of EHHT

In the paper by Hanchard et al. (2006), the authors conducted a comprehensive study on the power of popular detection tests to distinguish the selected from the non-selected alleles, including Hanchard's HS, Sabeti's EHH, Tajima's D test, Fu and Li's D test, Fay and Wu's H test, and Hudson's haplotype-partition method (Hudson et al., 1994; Hanchard et al., 2006; Sabeti et al., 2002; Tajima, 1989; Fu and Li, 1993; Fay and Wu, 2000). The simulation results suggested that Hanchard's HS and Sabeti's EHH are the two best tests. As a starting point, we can compare the performance of our EHHT with Hanchard's HS and Sabeti's EHH. Following a similar route, we perform coalescent simulations across different allele frequencies (of the core SNP) and recombination rates. Specifically, we generate 101 SNP sequences in a sample of 200 chromosomes that undergo a partial selected sweep with selection coefficient $s = 500$. The parameters involved in permutation include four recombination frequencies (ρ = 1.5, 3, 6, 9) and six frequencies of the derived alleles of the central SNP(0.1,

0.2, 0.4, 0.6, 0.8, 0.9). Each set of parameters is simulated 5,000 times so that a distribution of the test scores for the center SNP (i.e., SNP 50) is achieved. Power is calculated as the proportion of scores higher than the corresponding critical value from the standard normal distribution. For a fair comparison, we keep all parameters exactly the same as those in Hanchard's work except the number of SNPs. The reason is that the short length of SNPs will introduce bias in our tests due to the boundary truncation (Section B). Considering the scale of HapMap II data, our revision is reasonable. We also run a batch of simulation on 120 chromosomes to match the sample size from HapMap data.

From the results of Table V, we can see that at the nominal level $\alpha = 0.05$, the empirical power of the EHHT is higher than 0.92, irrespective of the recombination rate and allele frequency of the central SNP. Most of the EHHT empirical power levels are higher or close to 0.98. Comparing with Fig. 1 in Hanchard's paper, it clearly shows that the EHHT performs just as well as or even better than Hanchard's HS and Sabeti's EHH. Actually, the power of these two methods can be as low as 0.80 when the minor-allele frequency is 0.1 during the simulation. The rows in Table V marked by # contain results calculated using the same models and parameters as reported in Hanchard et al. (2006).

D. HapMap phase II data

After extensive simulation study, we apply the proposed score tests to the whole-genome SNP data of HapMap Phase II (The International HapMap Consortium, 2007). These data include genotypes for 3.1 million SNP from population samples of three continents: 60 CEPH Utah residents with ancestry from northern and western Europe (CEU), 60 Yoruba from Ibadan (YRI), Nigeria in Africa, and 45 Han Chinese

Table V. Power of the extended haplotype-based homozygosity test (EHHT). All results are based on $5,000$ simulations using software SelSim. The counterpart in Hanchard's simulation is marked with #.

Sample Size n	**Recomb. Rates ρ**	**Nominal Level α**	**Present Day Popu. Freq. of Derived Allele**					
			0.1	0.2	0.4	0.6	0.8	0.9
100	$1.5^{\#}$	0.05	0.9948	0.9944	0.9936	0.9930	0.9940	0.9776
		0.01	0.9918	0.9944	0.9936	0.9930	0.9916	0.9372
	$3^{\#}$	0.05	0.9836	0.9924	0.9876	0.9908	0.9894	0.9294
		0.01	0.9720	0.9924	0.9872	0.9908	0.9808	0.8628
	$6^{\#}$	0.05	0.9676	0.9904	0.9910	0.9906	0.9894	0.9276
		0.01	0.9342	0.9894	0.9910	0.9906	0.9822	0.8620
	9	0.05	0.9636	0.9900	0.9914	0.9904	0.9888	0.9236
		0.01	0.9220	0.9882	0.9914	0.9904	0.9804	0.8524
60	1.5	0.05	0.9604	0.9884	0.9828	0.9816	0.9638	0.8696
		0.01	0.9372	0.9874	0.9812	0.9796	0.9032	0.7776
	3	0.05	0.9350	0.9848	0.9810	0.9796	0.9238	0.7830
		0.01	0.8850	0.9778	0.9804	0.9780	0.8368	0.6928
	6	0.05	0.8930	0.9812	0.9810	0.9794	0.9368	0.7902
		0.01	0.8068	0.9648	0.9796	0.9778	0.8468	0.6860
	9	0.05	0.8524	0.9804	0.9816	0.9806	0.9288	0.7834
		0.01	0.7528	0.9550	0.9802	0.9790	0.8564	0.6770

from Beijing (CHB) and 45 Japanese from Tokyo (JPT) Japan of Asia. The two Asian samples are combined into one, CHB+JPT, as instructed by the HapMap Consortium. We use only the unrelated individuals from the three samples, omitting the children in the trio families from the CEU and YRI samples. The samples are available at (http://www.hapmap.org/downloads/phasing/2007-08_rel22/phased/).

E. Results in reported candidate regions

Due to the computational burdens in genome-wide scans, it is helpful to run our tests on short genome regions first, especially on those DNA segments reported with strong evidence of recent selection, which gives us a better view of the detection efficiency. As a major reference in our research, Sabeti et al. (2007) reported 20 autosomal candidate regions which show strong selection signals in at least one population. Note that our tests were designed for autosomal data. In 17 out of the 20 candidates, the EHHT scores show significant peaks for the selected population samples reported, hence,

there are extended stretches of homozygosity in these 17 regions and positive selection could be the driving force underneath. The three exceptions include a) a region around 78.3 Mb on chromosome 12, b) the *BCAS3* gene region on chromosome 17, c) the gene region of *CHST5, ADAT1* and *KARS* on chromosome 16. In the following paragraphs, we start our discussion from the candidate regions on chromosomes 2 and 15. In particular, we look at the regions containing the lactose tolerance gene *LCT* on chromosome 2 and the pigmentation gene *SLC24A5* on chromosome 15, which are widely studied for positive selection.

Figure 1 tells an interesting story of the *SLC24A5* gene on chromosome 15. The gene resides between the two dashed lines from 46.20 Mb to 46.22 Mb. The highest peak of EHHT occurs around 46.4 Mb, which was reported in Table 1 of Sabeti et al. (2007); the EHHT scores of CHB+JPT and YRI samples are very low and the test scores of YRI sample are uniformly the lowest. Our results are consistent with those of Sabeti et al. (2007) and Lamason et al. (2005), who argued for positive selection based on a striking reduction in heterozygosity in CEU sample.

In a 200 kb region around gene *HERC1* on chromosome 15, CHB+JPT sample shows signs of positive selection (Table 1, Sabeti et al., 2007). The EHHT scores are plotted in Figure 2. Again the gene is located between the dashed lines, from 61.69 Mb to 61.91 Mb. The EHHT scores of CHB+JPT are clearly highest within most part of *HERC1* gene. Hence, the CHB+JPT sample shows long extended haplotype homozygosity in the gene region.

The *LCT* gene sites between 136.26 Mb and 136.32 Mb on chromosome 2, and LD extends about 3.2 Mb around it in CEU sample (Bersaglieri et al., 2004; Enattah et al., 2002; Poulter, 2003). Two other genes are located in the same region, *RAB3GAP1* between 135.53 Mb and 135.64 Mb and *R3HDM1* between 136.01 Mb and 136.20 Mb. The EHHT scores in Figure 3 are noticeably higher in CEU sample

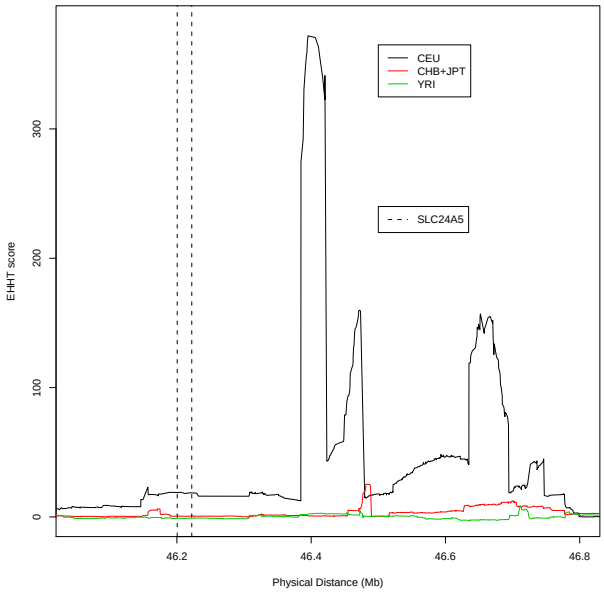

Fig. 1. The EHHT scores of three population samples in the region of *SLC24A5* on chromosome 15.

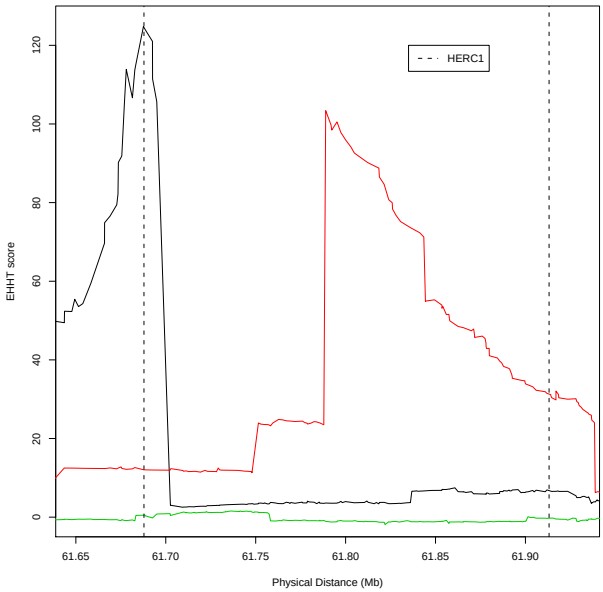

Fig. 2. The EHHT scores of three population samples in the region of *HERC1* on chromosome 15.

than in YRI and CHB+JPT samples, confirming previous results. Most striking of all is that the EHHT statistic spikes directly over the gene *R3HDM1* right next to gene *LCT*. Although this does not prove positive selection, the fact that a mutation deregulating the *LCT* gene occurs on the conserved haplotype strongly favors this interpretation. Because of the high density SNPs of HapMap data, high degree LD may not necessarily be the selection signal. Long extended haplotype homozygosity, however, can lead to high EHHT scores and interesting signals for further investigations.

Two other regions on chromosome 2, a 1.0 Mb region around the gene *EDAR* and an 800 kb region around 72.6 Mb, show strong evidence of selection in CHB+JPT sample (Table 1, Sabeti et al., 2007). The sharp EHHT peak for CHB+JPT sample locates very close to the *EDAR* region between 108.88 Mb and 108.97 Mb in Figure

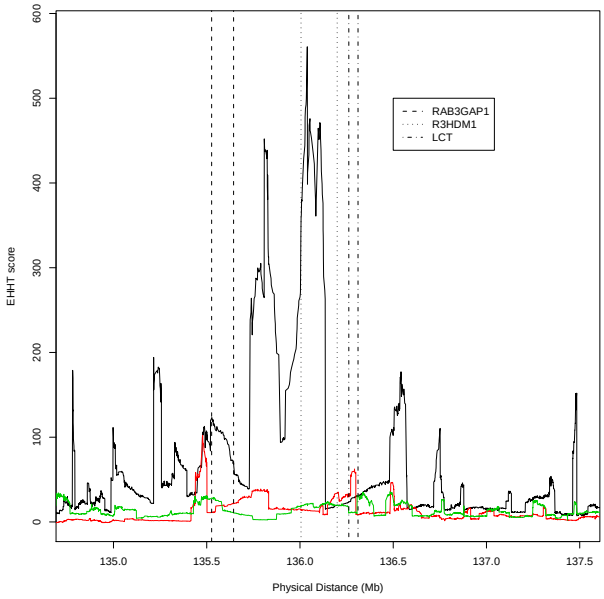

Fig. 3. The EHHT scores of three population samples in the region of *LCT* on chromosome 2.

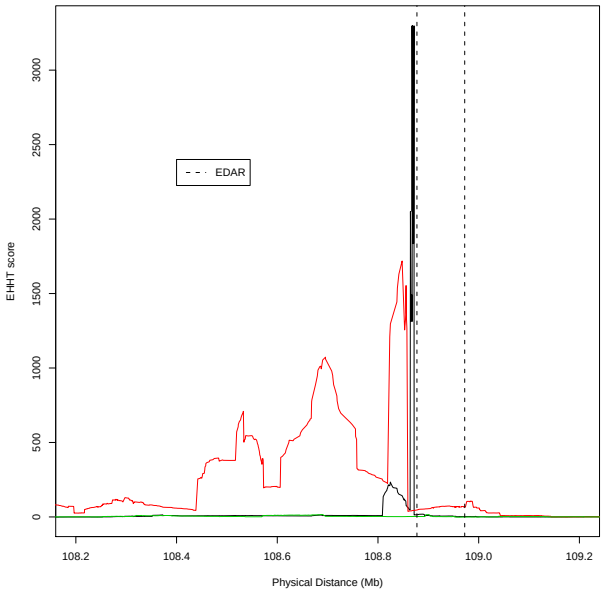

Fig. 4. The EHHT scores of three population samples in the region of *EDAR* on chromosome 15.

4. Also, our EHHT scores in Figure 5 confirm that the CHB+JPT sample has strong signal of selection in the 800 kb region around 72.6 Mb. Interestingly, the EHHT scores reach the highest in this region.

In a 1.2 Mb region around the gene *PDE11A*, both the CHB+JPT and CEU samples were reported to have strong signal of selection (Table 1, Sabeti et al. 2007). As we can see from Figure 6, the EHHT peaks of CHB+JPT and CEU samples show strong signals over the *PDE11A* region.

In a 300 kb region of the gene *SLC30A9* on chromosome 4, the CHB+JPT sample was reported to have strong signal of selection (Table 1, Sabeti et al. 2007). In Figure **7a**, we confirm the results by our EHHT scores because the sharp peak of CHB+JPT sample locates in the *SLC30A9* gene region, which is particular noteworthy.

In the region around 33.9 Mb on chromosome 4, all three samples of HapMap

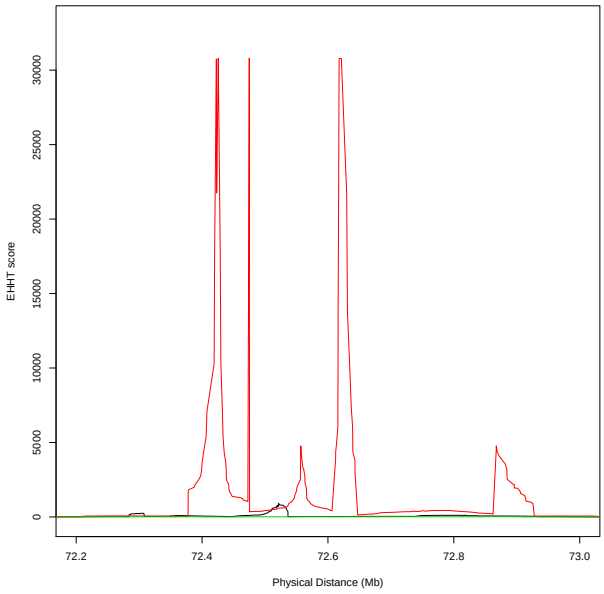

Fig. 5. The EHHT scores of three population samples around 72.6 Mb on chromosome 2.

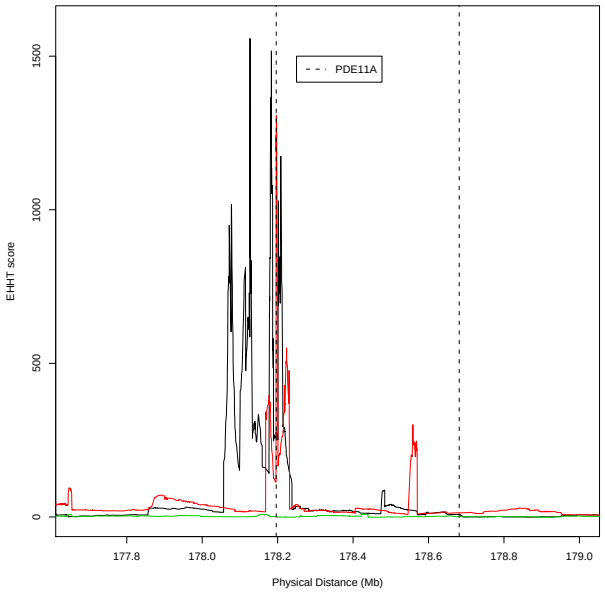

Fig. 6. The EHHT scores of three population samples in the region of *PDE11A* on chromosome 2.

Phase II data were reported to have selection signal (Table 1, Sabeti et al. 2007). Our EHHT scores confirm this for CHB+JPT and CEU samples in Figure 7**b**. Again in the region around 159 Mb on chromosome 4, the CHB+JPT sample was reported to provide strong signal of selection. We confirm the results in Figure 7**c**.

On chromosome 10, the CHB+JPT sample was reported to show strong signal of selection in a 400 kb region of the gene *PCDH15* (Table 1, Sabeti et al. 2007). We confirm the result by our EHHT scores in Figure 7**d**. Again on chromosome 10, the CEU and CHB+JPT samples showed strong signal of selection in a 300 kb region around 22.7 Mb (Table 1, Sabeti et al. 2007). We confirm the results by our EHHT scores in Figure 7**e**. Also on chromosome 10, the CEU sample was reported to show strong signal of selection in a 300 kb region around 3 Mb. We confirm the results by the EHHT scores in Figure 7**f**.

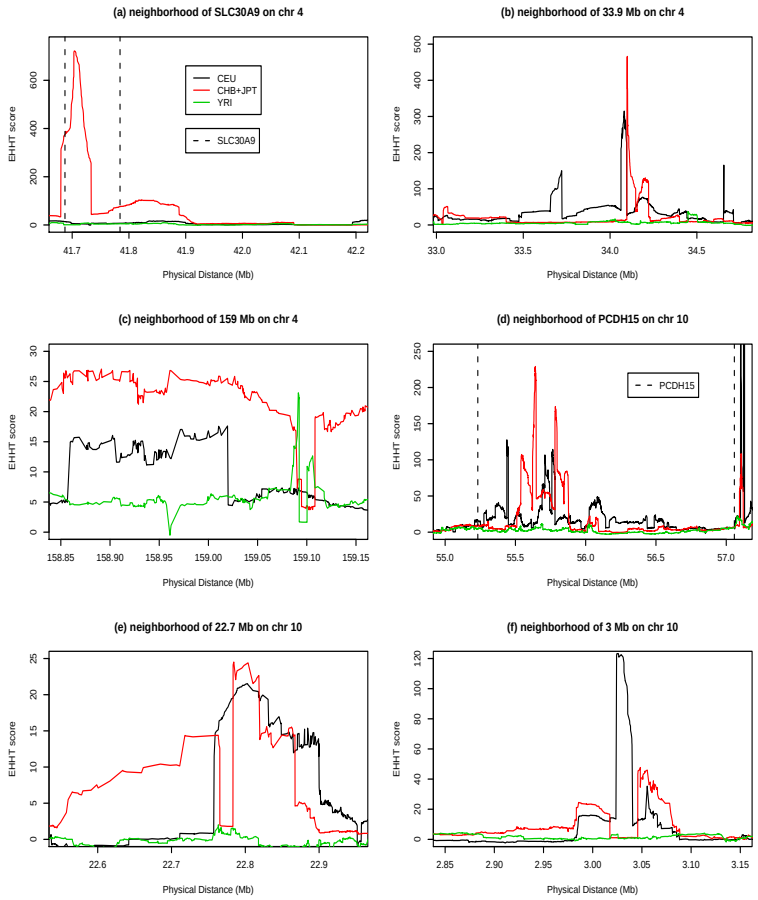

Fig. 7. The EHHT scores of three population samples of HapMap Phase II data in the candidate regions on chromosome 4 and 10. The dashed legend in Graph (**a**) indicated the location of gene *SLC30A9*, and similarly the location of *PCDH15* in Graph (**d**). **Abbreviation**: chr-chromosome.

On chromosome 1, there was strong signal of selection in a 400 kb region of the genes *BLZF1* and *SLC19A2* for CHB+JPT sample (Table 1, Sabeti et al., 2007). In Figure 8**a**, it is clear that the scores of the CHB+JPT sample are much higher than the scores of the CEU and YRI samples in the region of the two genes. A peak of EHHT occurs right between the two genes. Thus, our results confirm the findings of Sabeti et al. (2007). On chromosome 16, the CHB+JPT sample shows strong

selection signal in a region around 64.3 Mb. We confirm the result by high EHHT scores in Figure 8**b**. On chromosome 17, CHB+JPT sample has strong selection signal around 53.3 Mb. The EHHT scores of both CHB+JPT and CEU samples are high in Figure 8**c**. On chromosome 19, YRI sample was reported to have strong signal of selection around a region 43.5 Mb (Table 1, Sabeti et al. 2007). We confirm the result by the EHHT scores in Figure 8**d**. Again in the YRI sample, strong signal of selection was found in a 400 kb region that lay entirely within the gene *LARGE* on chromosome 22 (Table 1, Sabeti et al. 2007). We replicate this based on the EHHT score in Figure 8**e**.

On chromosome 12, the YRI sample showed strong signal of selection in a 800 kb region around 78.3 Mb (Table 1, Sabeti et al. 2007). We fail to confirm the results by the EHHT scores (Fig. 9**a**). In Table 1 of Sabeti et al. (2007), *BCAS3* on chromosome 17 was found to have strong signal of selection for the CEU sample. We fail to confirm the results (Fig. 9**b**). Actually, the EHHT scores of the CEU sample are lower than those of the CHB+JPT sample. In Table 1 of Sabeti et al. (2007), strong signal of selection was found in a 600 kb region of *CHST5*, *ADAT1*, and *KARS* on chromosome 16 for CHB+JPT and YRI samples. We fail to confirm the results by the EHHT scores (Fig. 9**c**).

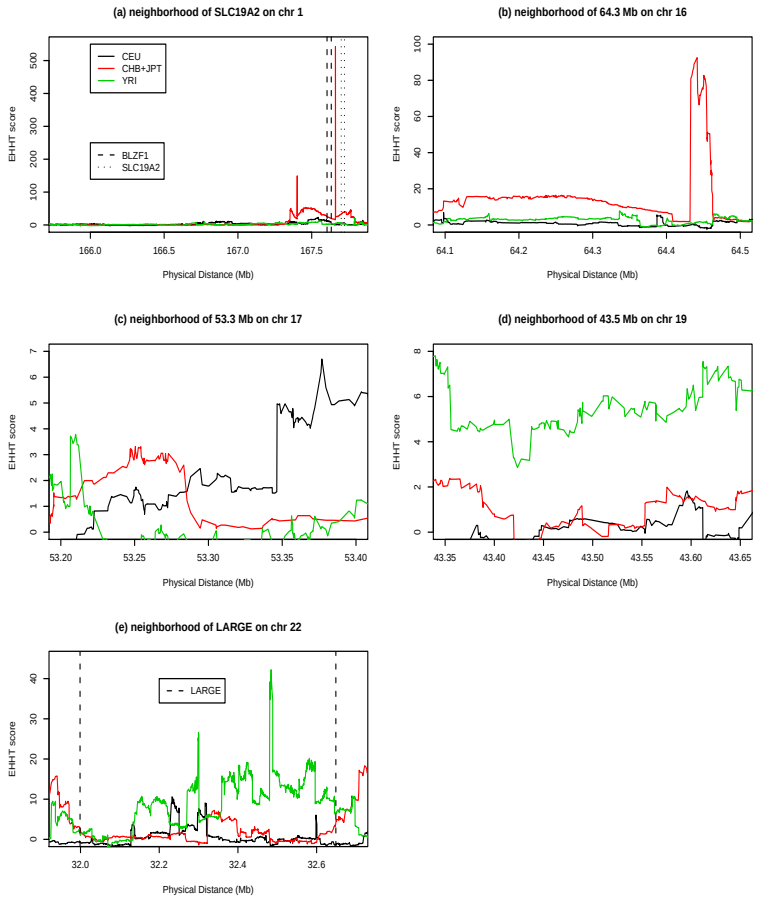

Fig. 8. The EHHT scores of three population sample of HapMap Phase II data in the candidate regions on chromosome 1, 16, 17, 19 and 22. In Graph (**a**), the dashed legend indicated the location of gene *BLZF1*, and the dotted legend indicated the location of gene *SLC19A2*. The dashed legend indicated the location of gene *LARGE* in Graph (**e**).

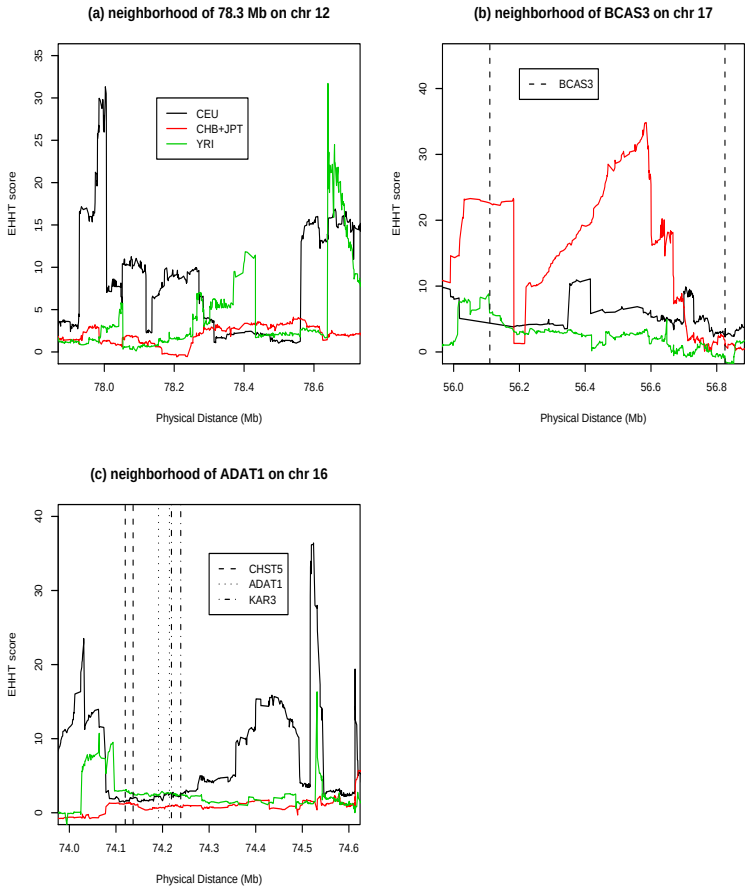

Fig. 9. The EHHT scores of three population sample of HapMap Phase II data in the candidate regions on chromosome 12, 16 and 17. The dashed legend indicated the location of gene *BCAS3* in Graph (**b**). In Graph (**c**), the dashed, dotted and dashed-dotted legends indicated the locations of *CHST5*, *ADAT1*, and *KARS* genes.

F. Impact of simple demographic models on type I error rates of EHHT

To investigate the impact of demographic population history on the EHHT, we perform coalescent simulations using ms Hudson (2002). We evaluate the type I error rates of EHHT under a few plausible population genetic demographic models, specif-

ically, four demographic models similar to those in Hanchard et al. (2006):

1. **Population structure**: two equal-sized sub-populations were simulated which exchanged migrants with a probability 0.1;

2. **Population expansion**: a rapid population growth was simulated with a current population size 10,000, and the population has a constant population size until 500 generations ago when it expanded exponentially by a factor of 100 to reach the current day population size;

3. **Population bottleneck 150/300**: a panmictic population was simulated which had a constant size 10,000 until $T_1 = 300$ generations ago when it underwent an instantaneous size reduction to 5,000, followed by a period 150 generations of constant size, and then followed by a rapid exponential population expansion in the last $T_2 = 150$ generations to reach a current day size 20,000.

4. **Population bottleneck 250/500**: a population similar to above, except $T_1 = 500$ and $T_2 = 250$.

Again, a genomic region of 101 SNPs is simulated with four recombination fractions $\rho = 1.5, 3, 6$ and 9. 5,000 samples are generated to calculate the empirical type I error rates. The results are reported in Table VI. When the recombination fraction ρ is 3, 6 or 9, the type I error rates are lower or around the nominal levels. Such as the results of Table IV, the type I error rates are generally higher than the nominal levels when $\rho = 1.5$. For the four models, the type I error rates of our EHHT are significantly lower than those of Hanchard's HS Hanchard et al. (2006). It suggests that the EHHT is reasonable robust and resistant to demographic change.

Table VI. Type I error rates of the extended haplotype-based homozygosity test (EHHT). All results are based on 5,000 simulations using software ms, and a genomic region of 101 SNPs is simulated.

Demographic Model	**Sample Size** n	ρ	**Nominal Level** α=0.05	α=0.01	**Sample Size** n	ρ	**Nominal Level** α=0.05	α=0.01
population structure	100	1.5	0.0512	0.0154	60	1.5	0.0538	0.0186
		3	0.0406	0.0082		3	0.0380	0.0118
		6	0.0292	0.0060		6	0.0300	0.0064
		9	0.0264	0.0046		9	0.0262	0.0060
population expansion	100	1.5	0.0664	0.0212	60	1.5	0.0514	0.0142
		3	0.0428	0.0094		3	0.0352	0.0090
		6	0.0414	0.0094		6	0.0272	0.0048
		9	0.0324	0.0058		9	0.0232	0.0042
population bottleneck 150/300	100	1.5	0.1040	0.0384	60	1.5	0.0954	0.0346
		3	0.0516	0.0180		3	0.0516	0.0170
		6	0.0366	0.0082		6	0.0362	0.0092
		9	0.0310	0.0066		9	0.0302	0.0068
population bottleneck 250/500	100	1.5	0.0864	0.0292	60	1.5	0.0870	0.0324
		3	0.0492	0.0140		3	0.0496	0.0154
		6	0.0348	0.0082		6	0.0324	0.0102
		9	0.0276	0.0072		9	0.0278	0.0084

G. New candidate regions for further investigation

Among the three proposed test statistics, EHHT is the most conservative. The high EHHT scores in a region indicate that there are extraordinary long stretches of homozygosity. In the 20 candidate regions reported previously, we find that EHHT scores show peaks in 17 of them. All these features encourage us to use EHHT in search of new candidate regions for further investigations.

We present 21 candidate regions and related SNPs in Tables VII — XXI for natural selection. To select a candidate SNP, we use four selection criteria as follows: 1) the selected SNP has high EHHT score of top one percentile, i.e., the EHHT score of the SNP is in the top one percentile of all SNPs of a chromosome in which the SNP is located, 2) the selected SNP has an allele which is likely to be newly derived by using the data from http://hg-wen.uchicago.edu/selection/frontpage.html of the University of Chicago (Voight et al., 2006), 3) the derived allele of the selected SNP has a high frequency which is larger than 0.5 in the tested population, 4) the derived allele of

the selected SNP is likely to be highly differentiated among the three populations of CHB+JPT, CEU, and YRI, in terms of the F_{st} score of the SNP is in the top one percentile of all F_{st} scores of SNPs on a chromosome (Weir and Cockerham, 1984; Akey et al., 2002, 2004). A candidate region is selected if there is a long segment of SNPs which satisfies the four criteria. In the 21 candidate regions, 3 are close to regions reported in Sabeti et al. (2007), and 12 are not reported; we count these 15 regions as new candidates. The remaining 6 regions are within regions reported in Sabeti et al. (2007). The region containing the least number of SNPs satisfying the criteria (7 SNPs) is located on chromosome 10, and it is reported in Table XVI because it is close to one candidate chr10:22.7 of Sabeti et al. (2007). Other regions contain 9 to 69 SNPs which satisfies the criteria.

In the Tables XXII and XXIII, we provide the maximum extended extended haplotype-based homozygosity test (EHHT) scores and related SNP information of the HapMap Phase II data of the three populations chromosome by chromosome. In the neighbor regions of these SNPs, there are usually long stretches of extended haplotype homozygosity and this could be an indication of positive selection. For instance, the Asian sample (CHB+JPT) was reported to have strong selection signal in the region of 72.6Mb on chromosome 2 (Sabeti et al. 2007). We confirm this finding by identifying two SNPs rs7594350 and rs1400582, which have highest EHHT score 30805.837 (Table XXII). The SNPs and their neighbor regions reported in the Tables XXII and XXIII could be useful for further investigation.

Based on the information of Tables XXII and XXIII, we then compare allele frequencies of the SNPs reported. Since selected alleles are likely to be highly differentiated between populations, we further select the SNPs whose allele frequency is significantly different among the three samples and then determine the selected population.

Table VII. Regions and SNPs which had highest EHHT scores, strongest differentiations, and high derived allele frequency in the tested population (> 0.5) on chromosome 1 and chromosome 2 of the HapMap Phase II data. * marked new regions; **abbreviations**: Chrms — Chromosome, Popu. — Population, Pct — percentile. In the first column, the **Genes** provided names and positions of genes which were located in a region.

Chrms, Tested Popu., Genes	**SNP Name**	**Position**	**Derived Allele Frequency**			F_{st} **Score**	**Pct of** F_{st}	**EHHT Score**	**Pct of EHHT**
			CEU	**CHB+JPT**	**YRI**				
Chrms 1, CHB+JPT*, *LHX8* (75,367-75,400kb), *SLC44A5* (75,440-75,849kb)	rs11162498	75329244	0.2167	0.8278	0.1167	0.5410	0.9925	94.979	0.9950
	rs11162513	75339512	0.2167	0.8333	0.1167	0.5480	0.9931	105.252	0.9955
	rs6671729	75342043	0.2167	0.8278	0.1167	0.5410	0.9925	106.522	0.9955
	rs12402007	75343282	0.2000	0.8333	0.1167	0.5582	0.9940	109.398	0.9957
	rs1007512	75344267	0.2250	0.8333	0.1167	0.5431	0.9927	109.894	0.9957
	rs17096272	75355512	0.2167	0.8333	0.1167	0.5480	0.9931	113.956	0.9960
	rs6663002	75360138	0.2167	0.8333	0.1167	0.5480	0.9931	113.721	0.9959
	rs10493552	75364816	0.2000	0.8333	0.1167	0.5582	0.9940	113.624	0.9959
	rs1526505	75371488	0.1917	0.8333	0.1167	0.5635	0.9944	112.276	0.9958
	rs12041465	75381637	0.1833	0.8333	0.1167	0.5689	0.9947	109.809	0.9957
	rs1144297	75512920	0.3750	0.8889	0.0833	0.5780	0.9953	71.381	0.9929
Chrms 2, CHB+JPT*	rs12614724	17200074	0.2667	0.9444	0.1750	0.6209	0.9967	842.262	0.9989
	rs1396074	17202963	0.2667	0.9444	0.1833	0.6148	0.9964	858.499	0.9989
	rs2135974	17203557	0.2667	0.9444	0.1833	0.6148	0.9964	810.547	0.9989
	rs11096723	17203869	0.2833	0.9444	0.1833	0.6059	0.9958	724.421	0.9987
	rs16983283	17217851	0.0333	0.7389	0.0833	0.6022	0.9955	156.695	0.9901
	rs7560778	17218020	0.0333	0.7389	0.0833	0.6022	0.9955	159.508	0.9903
	rs4832711	17218488	0.0333	0.7389	0.0833	0.6022	0.9955	176.744	0.9913
	rs1948984	17219024	0.0250	0.7389	0.0833	0.6102	0.9961	180.468	0.9915
	rs4832712	17219866	0.0417	0.7389	0.0917	0.5875	0.9941	215.376	0.9935
	rs925140	17226171	0.0917	0.9333	0.4000	0.6219	0.9968	247.588	0.9944
	rs287294	17235190	0.0833	0.7556	0.1083	0.5564	0.9909	263.261	0.9949
	rs12710631	17237305	0.0917	0.9333	0.4417	0.6126	0.9962	272.712	0.9952
	rs1507985	17240320	0.0917	0.9333	0.4417	0.6126	0.9962	280.998	0.9954
	rs2063164	17241504	0.0917	0.9333	0.4417	0.6126	0.9962	287.446	0.9955
	rs7563222	17242000	0.0917	0.9333	0.4500	0.6111	0.9962	291.606	0.9956
	rs1507977	17246940	0.1000	0.9333	0.4917	0.5975	0.9951	295.282	0.9957
	rs13010278	17247913	0.1000	0.9333	0.6500	0.6065	0.9958	292.050	0.9956
	rs1589272	17253381	0.1333	0.9389	0.7167	0.5971	0.9951	329.658	0.9963
	rs989555	17265872	0.0917	0.9333	0.4500	0.6111	0.9962	209.210	0.9933

Table VIII. One region and SNPs which had highest EHHT scores, strongest differentiations, and high derived allele frequency in the tested population (> 0.5) on chromosome 2 of the HapMap Phase II data.

Chromosome, Tested Popu.	**SNP Name**	**Position**	**Derived Allele Frequency**			F_{st} **Score**	**Pct of** F_{st}	**EHHT Score**	**Pct of EHHT**
			CEU	**CHB+JPT**	**YRI**				
Chrms 2, CHB+JPT	rs9309464	72387471	0.7917	0.9944	0.0083	0.8361	1.0000	1991.005	0.9996
	rs7595180	72389198	0.7833	0.9944	0.0083	0.8320	1.0000	2189.058	0.9996
	rs6752122	72399186	0.7917	0.9944	0.0167	0.8274	1.0000	3118.113	0.9997
	rs7558919	72402403	0.7917	0.9944	0.0750	0.7671	0.9997	4341.774	0.9998
	rs6546764	72406443	0.7917	0.9944	0.0667	0.7757	0.9997	5394.427	0.9998
	rs11126366	72407800	0.7917	0.9944	0.1583	0.6810	0.9987	7102.119	0.9999
	rs2420444	72419965	0.7917	0.9944	0.0083	0.8361	1.0000	17750.731	1.0000
	rs7594350	72425722	0.7917	0.9944	0.0083	0.8361	1.0000	30805.837	1.0000
	rs7558686	72432297	0.7917	0.9944	0.0667	0.7757	0.9997	7115.398	0.9999
	rs13390754	72437629	0.7917	0.9944	0.0750	0.7671	0.9997	3637.120	0.9998
	rs598496	72438429	0.7917	0.9944	0.0750	0.7671	0.9997	3123.944	0.9997
	rs598138	72438483	0.7917	0.9944	0.0667	0.7757	0.9997	2737.124	0.9997
	rs680495	72445746	0.7917	0.9944	0.0167	0.8274	1.0000	1568.588	0.9995
	rs590252	72461324	0.7917	0.9944	0.0667	0.7757	0.9997	1292.252	0.9993
	rs641939	72464683	0.7917	0.9944	0.0083	0.8361	1.0000	1220.543	0.9992
	rs630241	72475478	0.7917	0.9944	0.0583	0.7843	0.9998	334.550	0.9963
	rs628432	72477120	0.7917	0.9944	0.0167	0.8274	1.0000	360.481	0.9967
	rs659833	72496496	0.7917	0.9944	0.0083	0.8361	1.0000	390.677	0.9970
	rs3115351	72496883	0.7917	0.9944	0.0500	0.7930	0.9998	407.716	0.9972
	rs640610	72504445	0.7917	0.9944	0.0417	0.8016	0.9999	446.569	0.9976
	rs590345	72534217	0.6917	0.9944	0.1500	0.6504	0.9978	666.602	0.9986
	rs647242	72536445	0.7083	0.9944	0.1250	0.6807	0.9986	716.327	0.9987
	rs2203679	72542046	0.7083	0.9944	0.0833	0.7234	0.9993	1013.560	0.9991
	rs653220	72561382	0.7083	0.9944	0.0083	0.8010	0.9999	3219.831	0.9997
	rs6714595	72565756	0.7083	0.9944	0.0083	0.8010	0.9999	1919.837	0.9996
	rs11686713	72662046	0.7083	0.9944	0.0417	0.7664	0.9996	178.415	0.9914
	rs11677707	72732262	0.7083	0.9944	0.0250	0.7837	0.9998	362.720	0.9967
	rs6724529	72734249	0.7083	0.9944	0.0250	0.7837	0.9998	375.546	0.9969
	rs4852886	72736040	0.7083	0.9944	0.0917	0.7148	0.9992	388.213	0.9970
	rs11685114	72749123	0.7083	0.9944	0.0083	0.8010	0.9999	408.743	0.9973
	rs4852891	72798521	0.7250	0.9944	0.0167	0.7983	0.9999	405.396	0.9972
	rs7588400	72830661	0.7250	0.9944	0.0167	0.7983	0.9999	306.541	0.9959
	rs970577	72859087	0.8000	0.9389	0.2000	0.5623	0.9917	225.758	0.9938
	rs1876490	72905859	0.6833	0.9333	0.0167	0.7049	0.9990	1588.944	0.9995

Table IX. One region and SNPs which had highest EHHT scores, strongest differentiations, and high derived allele frequency in the tested population (> 0.5) on chromosome 2 of the HapMap Phase II data. * marked new regions; **abbreviations**: Chrms — Chromosome, Popu. — Population, Pct — percentile.

Chromosome, Tested Popu.	**SNP Name**	**Position**	**Derived Allele Frequency**			F_{st} **Score**	**Pct of** F_{st}	**EHHT Score**	**Pct of EHHT**
			CEU	**CHB+JPT**	**YRI**				
Chrms 2, CHB+JPT*	rs1514999	103484100	0.4750	0.9333	0.1000	0.5993	0.9953	164.562	0.9907
	rs7595711	103519644	0.4750	0.9333	0.1000	0.5993	0.9953	588.340	0.9983
	rs6734144	103521977	0.4750	0.9333	0.1000	0.5993	0.9953	626.421	0.9985
	rs4851029	103526217	0.4500	0.9333	0.1167	0.5873	0.9941	639.747	0.9985
	rs6721892	103529619	0.4583	0.9444	0.1250	0.5933	0.9948	640.805	0.9985
	rs1983305	103530871	0.4583	0.9444	0.1250	0.5933	0.9948	601.374	0.9984
	rs4851031	103531382	0.4583	0.9444	0.1250	0.5933	0.9948	576.552	0.9983
	rs1451974	103540096	0.4583	0.9444	0.1333	0.5856	0.9939	363.404	0.9967
	rs7580027	103544809	0.4583	0.9389	0.1250	0.5856	0.9939	269.354	0.9951
	rs6543217	103549697	0.4583	0.9444	0.1250	0.5933	0.9948	261.158	0.9948
	rs6543218	103549969	0.4583	0.9444	0.1250	0.5933	0.9948	251.652	0.9945
	rs13411937	103553084	0.4583	0.9444	0.1250	0.5933	0.9948	234.368	0.9941
	rs7565635	103558332	0.4583	0.9444	0.1417	0.5779	0.9931	187.937	0.9919
	rs6738539	103562187	0.4583	0.9444	0.1417	0.5779	0.9931	172.925	0.9911
	rs7591265	103562871	0.4583	0.9444	0.1250	0.5933	0.9948	168.412	0.9909
	rs6727525	103569144	0.4583	0.9444	0.1250	0.5933	0.9948	275.921	0.9953
	rs1869073	103570952	0.4583	0.9444	0.1417	0.5779	0.9931	327.638	0.9962
	rs10188273	103575435	0.3917	0.9444	0.1250	0.6091	0.9960	384.636	0.9970
	rs4851660	103584221	0.4583	0.9444	0.1417	0.5779	0.9931	441.440	0.9976
	rs10189533	103585672	0.4583	0.9444	0.1333	0.5856	0.9939	488.319	0.9979
	rs10192716	103586576	0.4583	0.9444	0.1250	0.5933	0.9948	546.241	0.9982
	rs1451988	103589021	0.4583	0.9444	0.1250	0.5933	0.9948	659.861	0.9986
	rs6543223	103594301	0.4583	0.9444	0.1250	0.5933	0.9948	886.581	0.9989
	rs13396809	103595295	0.4667	0.9444	0.1250	0.5919	0.9945	1067.541	0.9991
	rs7570362	103597727	0.4583	0.9444	0.1250	0.5933	0.9948	1283.435	0.9993
	rs12328095	103603254	0.4583	0.9389	0.1417	0.5702	0.9923	1526.657	0.9994
	rs6719978	103604933	0.4583	0.9444	0.1333	0.5856	0.9939	1379.914	0.9994
	rs1584705	103606168	0.4583	0.9444	0.1417	0.5779	0.9931	1286.076	0.9993
	rs11123965	103606637	0.4583	0.9444	0.1250	0.5933	0.9948	1188.888	0.9992
	rs11888473	103606852	0.4583	0.9444	0.1417	0.5779	0.9931	1091.811	0.9992

Table X. Regions and SNPs which had highest EHHT scores, strongest differentiations, and high derived allele frequency in the tested population (> 0.5) on chromosome 2 and chromosome 3 of the HapMap Phase II data. * marked new regions; **abbreviations**: Chrms — Chromosome, Popu. — Population, Pct — percentile. In the first column, the **Genes** provided names and positions of genes which were located in a region.

Chromosome, Tested Popu.	**SNP Name**	**Position**	**Derived Allele Frequency**			F_{st} **Score**	**Pct of** F_{st}	**EHHT Score**	**Pct of EHHT**
			CEU	**CHB+JPT**	**YRI**				
Chrms 2,	rs10179602	108451805	0.3833	0.8722	0.0250	0.6124	0.9962	290.676	0.9956
CHB+JPT	rs1053027	108457987	0.3833	0.8778	0.0917	0.5534	0.9905	349.171	0.9965
LIMS1	rs2077472	108492518	0.3833	0.8944	0.1000	0.5670	0.9921	380.171	0.9969
(108,571-108,670kb)	rs12613554	108527107	0.3917	0.8722	0.0167	0.6190	0.9966	651.248	0.9985
	rs12473539	108531804	0.3917	0.8722	0.0833	0.5520	0.9903	696.403	0.9987
	rs1469965	108536852	0.3917	0.8722	0.0833	0.5520	0.9903	536.393	0.9981
	rs10179040	108552213	0.3917	0.8722	0.0833	0.5520	0.9903	529.817	0.9981
	rs11123708	108562291	0.3917	0.8722	0.0833	0.5520	0.9903	471.995	0.9978
	rs10187016	108563195	0.3833	0.8778	0.0833	0.5613	0.9915	461.466	0.9977
	rs13413437	108564962	0.3917	0.8722	0.0833	0.5520	0.9903	427.252	0.9974
	rs13422997	108572896	0.3917	0.8722	0.0833	0.5520	0.9903	195.172	0.9923
	rs7565372	108587342	0.3917	0.8722	0.0833	0.5520	0.9903	200.544	0.9927
	rs6707379	108600999	0.3917	0.8722	0.0750	0.5601	0.9914	199.593	0.9926
	rs12469016	108622967	0.3917	0.8722	0.0750	0.5601	0.9914	500.032	0.9979
Chrms, 3	rs9827968	106178646	0.4250	0.8556	0.0667	0.5404	0.9915	119.307	0.9931
CHB+JPT*	rs1503079	106241722	0.4917	0.9278	0.1333	0.5577	0.9933	230.958	0.9976
	rs1566718	106244327	0.4917	0.9278	0.1417	0.5497	0.9926	226.533	0.9975
	rs1566717	106250199	0.4917	0.9278	0.1417	0.5497	0.9925	203.063	0.9968
	rs13090983	106250823	0.4917	0.9278	0.1417	0.5497	0.9926	193.690	0.9967
	rs12633740	106252196	0.4917	0.9278	0.1417	0.5497	0.9926	175.793	0.9962
	rs12637494	106256655	0.4917	0.9278	0.1417	0.5497	0.9925	162.694	0.9956
	rs10933802	106256994	0.4917	0.9278	0.1417	0.5497	0.9925	155.645	0.9952
	rs10933803	106257871	0.4917	0.9278	0.0500	0.6401	0.9981	145.816	0.9947
	rs2047806	106261500	0.5250	0.9778	0.1500	0.6091	0.9965	122.116	0.9933
	rs2134526	106262934	0.4917	0.9278	0.1417	0.5497	0.9925	114.486	0.9927
	rs6800325	106265702	0.5417	0.9778	0.1417	0.6161	0.9969	1473.750	0.9999
	rs10933807	106267020	0.4917	0.9278	0.1417	0.5497	0.9925	1006.902	0.9997
	rs2895296	106271452	0.4917	0.9278	0.1167	0.5737	0.9946	571.559	0.9993
	rs1503084	106274502	0.4917	0.9278	0.1167	0.5737	0.9946	96.787	0.9912
	rs10933809	106283040	0.4917	0.9278	0.1167	0.5737	0.9946	121.231	0.9932
	rs1503085	106284786	0.4917	0.9278	0.1167	0.5737	0.9946	125.271	0.9934
	rs1503075	106289543	0.4167	0.9056	0.0667	0.6051	0.9963	156.038	0.9953
	rs1503158	106297795	0.3917	0.9056	0.0667	0.6110	0.9966	167.596	0.9958
	rs12492439	106299176	0.4000	0.9056	0.0333	0.6418	0.9982	175.354	0.9962
	rs870279	106305849	0.4000	0.8889	0.0750	0.5793	0.9951	174.778	0.9962
	rs12492301	106306013	0.4000	0.8889	0.0667	0.5874	0.9956	177.173	0.9963

Table XI. Two regions and SNPs which had highest EHHT scores, strongest differentiations, and high derived allele frequency in the tested population (> 0.5) on chromosome 4 of the HapMap Phase II data. * marked new regions; **abbreviations**: Chrms — Chromosome, Popu. — Population, Pct — percentile. # marked region which was close to a candidate region found in Table 1, Sabeti et al. (2007).

Chromosome, Tested Popu.	**SNP Name**	**Position**	**Derived Allele Frequency**			F_{st} **Score**	**Pct of** F_{st}	**EHHT Score**	**Pct of EHHT**
			CEU	**CHB+JPT**	**YRI**				
Chrms 4, CEU	rs10002222	33667869	0.9000	0.6778	0.0250	0.5912	0.9942	89.357	0.9941
	rs12645236	33691462	0.8667	0.6833	0.0250	0.5671	0.9924	98.490	0.9948
	rs6821548	33706776	0.9000	0.6833	0.0250	0.5940	0.9944	129.362	0.9965
	rs7665973	34073859	0.8917	0.9833	0.3333	0.5464	0.9905	259.032	0.9990
	rs10031884	34074934	0.8833	0.9833	0.2417	0.6353	0.9970	278.189	0.9991
	rs6830615	34077763	0.8833	0.9833	0.3083	0.5660	0.9923	285.510	0.9992
	rs1364911	34078305	0.8833	0.9833	0.2167	0.6612	0.9981	290.995	0.9992
	rs6817183	34174660	0.8333	0.8167	0.0250	0.6400	0.9973	70.715	0.9917
	rs1842531	34179397	0.8333	0.8111	0.0250	0.6350	0.9970	71.550	0.9918
	rs6847787	34183722	0.9333	0.8278	0.1333	0.6037	0.9949	75.518	0.9927
	rs6812860	34185807	0.9750	0.9722	0.1417	0.8008	1.0000	76.661	0.9929
	rs10033586	34190319	0.9000	0.8167	0.1250	0.5752	0.9930	75.174	0.9926
	rs4859322	34198313	0.8333	0.8111	0.0250	0.6350	0.9970	72.822	0.9921
	rs4859283	34203697	0.8333	0.8111	0.0250	0.6350	0.9970	72.773	0.9921
Chrms 4, CHB+JPT*,#	rs714226	41521093	0.9333	0.9944	0.2417	0.6948	0.9988	58.060	0.9940
	rs6840961	41522632	0.8417	0.9833	0.2250	0.6237	0.9962	57.797	0.9940
	rs9991121	41523790	0.8500	0.9778	0.1250	0.7251	0.9993	54.141	0.9934
	rs4623048	41528244	0.7583	0.9611	0.0167	0.7677	0.9999	53.418	0.9931
	rs9998239	41529367	0.9083	0.9722	0.2083	0.6728	0.9983	52.918	0.9930
	rs6839376	41544538	0.8000	0.9611	0.1000	0.6990	0.9990	50.703	0.9925
	rs6447118	41550330	0.8000	0.9611	0.1000	0.6990	0.9990	49.738	0.9922
	rs7356183	41554214	0.8083	0.9611	0.1000	0.7035	0.9991	49.412	0.9921
	rs7660832	41584424	0.7833	0.9611	0.2417	0.5411	0.9900	45.336	0.9907
	rs2660335	41680446	0.7583	0.9667	0.0917	0.6951	0.9989	316.060	0.9991
	rs2581435	41685047	0.7583	0.9667	0.0917	0.6951	0.9989	367.830	0.9993
	rs4540084	41696545	0.7583	0.9667	0.0917	0.6951	0.9989	403.192	0.9994
	rs2660331	41696942	0.7583	0.9667	0.0917	0.6951	0.9989	415.118	0.9994
	rs2581455	41697890	0.7333	0.9667	0.0583	0.7201	0.9993	442.977	0.9994
	rs2660330	41698140	0.7667	0.9667	0.0917	0.6989	0.9989	470.647	0.9995
	rs1047626	41698428	0.7333	0.9667	0.0583	0.7201	0.9993	484.840	0.9995
	rs2581453	41698592	0.7583	0.9722	0.0917	0.7027	0.9990	500.064	0.9995
	rs2660329	41700744	0.7583	0.9667	0.1500	0.6336	0.9969	528.134	0.9995
	rs2660326	41701828	0.7583	0.9667	0.1500	0.6336	0.9968	567.847	0.9996
	rs2660325	41701960	0.7583	0.9667	0.1500	0.6336	0.9969	579.248	0.9996
	rs2581449	41701981	0.7583	0.9667	0.1500	0.6336	0.9969	589.284	0.9996
	rs2581448	41702151	0.7583	0.9667	0.1500	0.6336	0.9969	604.229	0.9996
	rs2660323	41702769	0.7583	0.9667	0.1500	0.6336	0.9969	613.770	0.9996
	rs9884564	41703000	0.7333	0.9722	0.0333	0.7540	0.9997	714.736	0.9997
	rs2581443	41703898	0.7583	0.9667	0.1500	0.6336	0.9969	721.930	0.9998
	rs2660319	41705641	0.7333	0.9667	0.1000	0.6760	0.9985	719.208	0.9998
	rs2581441	41706727	0.7333	0.9667	0.0333	0.7467	0.9996	704.512	0.9997
	rs2581426	41712257	0.7333	0.9667	0.0333	0.7467	0.9996	664.020	0.9997
	rs2581424	41713389	0.7583	0.9667	0.1500	0.6336	0.9969	630.272	0.9997

Table XII. Continuation of Table XI. In the first column, the **Genes** provided names and positions of genes which were located in a region.

Chromosome, Tested Popu.	**SNP Name**	**Position**	**Derived Allele Frequency**			F_{st} **Score**	**Pct of** F_{st}	**EHHT Score**	**Pct of EHHT**
			CEU	**CHB+JPT**	**YRI**				
Chrms 4,	rs1848180	41713701	0.7333	0.9667	0.0333	0.7467	0.9996	612.026	0.9996
CHB+JPT	rs10805092	41714239	0.7583	0.9667	0.1500	0.6336	0.9968	593.444	0.9996
CCDC4	rs2581420	41716994	0.7583	0.9667	0.1500	0.6336	0.9968	525.486	0.9995
(41,808-41,850kb),	rs10461065	41719050	0.7583	0.9667	0.1500	0.6336	0.9969	477.692	0.9995
TMEM33	rs10461059	41720699	0.7583	0.9667	0.1500	0.6336	0.9968	421.180	0.9994
(41,632-41,653kb)	rs9998823	41720912	0.7583	0.9722	0.1500	0.6415	0.9973	408.751	0.9994
WDR21B	rs4241695	41722202	0.7583	0.9667	0.1500	0.6336	0.9969	396.399	0.9994
(41,679-41,680kb)	rs9654067	41723554	0.7583	0.9667	0.1500	0.6336	0.9968	373.813	0.9993
	rs3827588	41725973	0.7583	0.9667	0.1500	0.6336	0.9968	343.577	0.9992
	rs3827590	41726194	0.7083	0.9722	0.1500	0.6228	0.9961	334.513	0.9992
	rs3827591	41726229	0.7083	0.9667	0.1500	0.6149	0.9957	325.588	0.9992
	rs4861155	41726781	0.7333	0.9667	0.0333	0.7467	0.9996	309.138	0.9991
	rs7683204	41728350	0.7583	0.9667	0.1500	0.6336	0.9968	294.067	0.9991
	rs11725865	41731391	0.7667	0.9667	0.1417	0.6460	0.9975	273.820	0.9990
	rs10006383	41733603	0.7667	0.9667	0.1417	0.6460	0.9976	43.632	0.9901
	rs12507609	41736854	0.7667	0.9667	0.1500	0.6373	0.9972	44.266	0.9904
	rs4377621	41738061	0.7083	0.9667	0.0333	0.7375	0.9994	44.915	0.9906
	rs10433708	41745215	0.7667	0.9667	0.1417	0.6460	0.9976	45.227	0.9907
	rs10938170	41750639	0.7667	0.9667	0.1417	0.6460	0.9976	47.358	0.9914
	rs3804192	41761196	0.7667	0.9667	0.1417	0.6460	0.9976	49.648	0.9922
	rs12647092	41766976	0.7083	0.9667	0.0333	0.7375	0.9994	59.678	0.9943
	rs10019356	41768229	0.7667	0.9667	0.1500	0.6373	0.9972	60.754	0.9946
	rs4438791	41769322	0.7083	0.9667	0.0333	0.7375	0.9994	61.825	0.9947
	rs7660223	41775871	0.7667	0.9667	0.1000	0.6900	0.9988	71.182	0.9957
	rs11725543	41781944	0.7667	0.9667	0.1417	0.6460	0.9975	73.344	0.9959
	rs10002107	41782303	0.7667	0.9667	0.1417	0.6460	0.9975	74.775	0.9960
	rs12511999	41786014	0.7083	0.9667	0.0333	0.7375	0.9994	76.895	0.9962
	rs6832890	41801295	0.7583	0.9667	0.0833	0.7040	0.9991	83.865	0.9966
	rs6447128	41801872	0.7500	0.9778	0.0833	0.7156	0.9992	86.149	0.9968
	rs7682049	41807491	0.6750	0.9500	0.0167	0.7234	0.9993	89.610	0.997
	rs13756	41807998	0.6750	0.9500	0.0167	0.7234	0.9993	90.879	0.9971
	rs16854014	41812231	0.9333	0.9833	0.3667	0.5479	0.9907	100.079	0.9976
	rs2880666	41815266	0.6750	0.9500	0.0167	0.7234	0.9993	98.591	0.9975
	rs6447131	41824754	0.6750	0.9556	0.0500	0.6949	0.9988	101.828	0.9977
	rs6447132	41828823	0.6583	0.9556	0.0583	0.6817	0.9986	101.778	0.9977
	rs7664565	41829776	0.6667	0.9556	0.0583	0.6838	0.9987	101.689	0.9977
	rs6848386	41841414	0.6750	0.9556	0.1500	0.5900	0.9941	101.398	0.9977
	rs6856819	41844970	0.6917	0.9556	0.2000	0.5430	0.9903	99.008	0.9975
	rs4861024	41847307	0.6917	0.9556	0.1500	0.5944	0.9944	98.502	0.9970
	rs4449446	41849931	0.6833	0.9500	0.1583	0.5758	0.9931	96.958	0.9970

Table XIII. One region and SNPs which had highest EHHT scores, strongest differentiations, and high derived allele frequency in the tested population (> 0.5) on chromosome 5 of the HapMap Phase II data. * marked new regions; **abbreviations**: Chrms — Chromosome, Popu. — Population, Pct — percentile.

Chromosome, Tested Popu.	**SNP Name**	**Position**	**Derived Allele Frequency**			F_{st} **Score**	**Pct of** F_{st}	**EHHT Score**	**Pct of EHHT**
			CEU	**CHB+JPT**	**YRI**				
Chrms 5, CHB+JPT*	rs397317	117006587	0.6583	0.9722	0.0250	0.7391	0.9996	92.618	0.9948
	rs267035	117010380	0.6583	0.9722	0.0250	0.7391	0.9996	112.143	0.9964
	rs267033	117010655	0.6583	0.9722	0.0250	0.7391	0.9996	107.010	0.9960
	rs267030	117011477	0.6583	0.9722	0.0250	0.7391	0.9996	119.091	0.9969
	rs192378	117011538	0.6583	0.9722	0.0250	0.7391	0.9996	121.734	0.9970
	rs197864	117014844	0.6583	0.9722	0.0250	0.7391	0.9996	126.235	0.9971
	rs842002	117023413	0.6583	0.9722	0.0250	0.7391	0.9996	92.058	0.9948
	rs842003	117023684	0.6583	0.9722	0.0250	0.7391	0.9996	93.725	0.9949
	rs1686409	117030985	0.6583	0.9722	0.0250	0.7391	0.9996	96.408	0.9952
	rs1686410	117030996	0.6583	0.9722	0.0250	0.7391	0.9996	94.739	0.9950
	rs10036241	117033421	0.7083	0.9667	0.1500	0.6149	0.9968	87.005	0.9943
	rs2416472	117033845	0.6917	0.9667	0.0417	0.7233	0.9992	84.739	0.9941
	rs7714451	117037818	0.8333	0.9667	0.1750	0.6463	0.9979	75.177	0.9927
	rs7724328	117515965	0.4083	0.9778	0.0917	0.6812	0.9988	773.232	0.9998
	rs4317366	117516317	0.4083	0.9778	0.0167	0.7513	0.9996	713.005	0.9998
	rs6872244	117517311	0.4083	0.9667	0.0167	0.7364	0.9993	644.770	0.9997
	rs1479207	117520698	0.4083	0.9778	0.0167	0.7513	0.9996	97.216	0.9953
	rs10079352	117522539	0.4083	0.9778	0.0167	0.7513	0.9996	103.815	0.9958
	rs4639272	117524321	0.4167	0.9778	0.1167	0.6568	0.9982	109.814	0.9963
	rs4401605	117524359	0.4167	0.9778	0.0167	0.7496	0.9996	110.889	0.9963
	rs7721999	117524545	0.4250	0.9778	0.0167	0.7481	0.9996	111.909	0.9964
	rs2900117	117525764	0.5000	0.9778	0.0333	0.7231	0.9992	116.020	0.9967
	rs734155	117531602	0.4917	0.9778	0.1500	0.6125	0.9967	161.408	0.9980
	rs13356156	117533769	0.4583	0.9778	0.0583	0.7027	0.9990	164.766	0.9981
	rs1479196	117534564	0.4583	0.9778	0.0583	0.7027	0.9990	168.118	0.9981
	rs1382721	117543368	0.5000	0.9778	0.0500	0.7066	0.9990	187.399	0.9984
	rs1382720	117547386	0.4583	0.9778	0.0500	0.7107	0.9991	199.031	0.9987
	rs6883098	117588162	0.5000	0.9778	0.2167	0.5522	0.9925	192.883	0.9985
	rs1479225	117588467	0.5000	0.9778	0.2083	0.5594	0.9935	192.319	0.9985
	rs6859099	117590938	0.4750	0.9722	0.2167	0.5486	0.9922	187.462	0.9985
	rs1871367	117599244	0.4167	0.9667	0.0167	0.7347	0.9993	169.643	0.9982
	rs7341174	117620240	0.5833	0.9667	0.2167	0.5287	0.9900	139.264	0.9974

Table XIV. One region and SNPs which had highest EHHT scores, strongest differentiations, and high derived allele frequency in the tested population (> 0.5) on chromosome 7 of the HapMap Phase II data. * marked new regions; **abbreviations**: Chrms — Chromosome, Popu. — Population, Pct — percentile.

Chromosome, Tested Popu.	**SNP Name**	**Position**	**Derived Allele Frequency**			F_{st} **Score**	**Pct of** F_{st}	**EHHT Score**	**Pct of EHHT**
			CEU	**CHB+JPT**	**YRI**				
Chrms 7, CEU*	rs7789561	119168428	0.7667	0.9611	0.1083	0.6736	0.9987	82.129	0.9962
	rs11978043	119170356	0.7667	0.9611	0.1083	0.6736	0.9987	85.747	0.9965
	rs1404083	119187663	0.7667	0.9611	0.1083	0.6736	0.9987	96.623	0.9973
	rs12706259	119197681	0.7667	0.9556	0.1083	0.6661	0.9984	90.741	0.9968
	rs13239182	119212898	0.7667	0.9611	0.0667	0.7180	0.9995	90.151	0.9967
	rs12536246	119216218	0.7667	0.9611	0.0667	0.7180	0.9995	89.162	0.9967
	rs6466713	119221432	0.7667	0.9611	0.1083	0.6736	0.9987	82.437	0.9963
	rs940412	119228681	0.7667	0.9611	0.1083	0.6736	0.9987	2036.208	1.0000
	rs1916859	119230892	0.7667	0.9611	0.1083	0.6736	0.9987	920.089	0.9999

Table XV. Two regions and SNPs which had highest EHHT scores, strongest differentiations, and high derived allele frequency in the tested population (> 0.5) on chromosome 8 of the HapMap Phase II data. * marked new regions; **abbreviations**: Chrms — Chromosome, Popu. — Population, Pct — percentile. In the first column, the **Genes** provided names and positions of genes which were located in a region.

Chromosome, Tested Popu.	**SNP Name**	**Position**	**Derived Allele Frequency**			F_{st} **Score**	**Pct of** F_{st}	**EHHT Score**	**Pct of EHHT**
			CEU	**CHB+JPT**	**YRI**				
Chrms 8, CEU*	rs6989187	50635491	0.8417	0.8667	0.1333	0.5726	0.9947	172.591	0.9949
	rs9643617	50652967	0.8417	0.8667	0.1250	0.5819	0.9953	211.534	0.9965
	rs9643406	50670255	0.8417	0.8667	0.1250	0.5819	0.9953	283.270	0.9984
	rs6992847	50671501	0.8417	0.8667	0.1250	0.5819	0.9953	322.309	0.9989
	rs1352112	50935021	0.8417	0.8667	0.1667	0.5359	0.9909	119.764	0.9903
	rs9886451	50935477	0.8417	0.8667	0.1667	0.5359	0.9909	122.549	0.9906
	rs1552372	50935855	0.8417	0.8667	0.1667	0.5359	0.9909	125.420	0.9909
	rs9298228	50937171	0.8417	0.8667	0.1167	0.5911	0.9959	128.378	0.9912
	rs9298229	50937251	0.8417	0.8667	0.1667	0.5359	0.9909	131.424	0.9915
	rs10089399	50938876	0.8333	0.8667	0.1667	0.5305	0.9901	137.789	0.9920
	rs10092384	50939034	0.8417	0.8667	0.1667	0.5359	0.9909	141.128	0.9923
	rs11785147	50940355	0.8417	0.8667	0.1667	0.5359	0.9909	152.275	0.9934
	rs13251355	50943725	0.9833	0.8667	0.2583	0.5550	0.9935	173.889	0.9949
Chrms 8, CEU* *PCMTD1* (52,893-52.936kb)	rs10464943	52876153	0.9083	0.5722	0.0500	0.5327	0.9905	655.337	0.9998
	rs756484	52877010	0.9083	0.5722	0.0417	0.5417	0.9921	651.099	0.9998
	rs10958298	52882635	0.9083	0.5722	0.0417	0.5417	0.9921	216.268	0.9967
	rs13275235	52887958	0.9083	0.5722	0.0417	0.5417	0.9918	282.938	0.9984
	rs4873608	52890882	0.9083	0.5722	0.0417	0.5417	0.9918	290.769	0.9984
	rs13257067	52898125	0.9083	0.5722	0.0417	0.5417	0.9918	312.322	0.9988
	rs753539	52904315	0.9083	0.5722	0.0417	0.5417	0.9921	329.594	0.9989
	rs10504128	52911112	0.9083	0.5722	0.0417	0.5417	0.9918	467.469	0.9995
	rs5002416	52913710	0.9083	0.5722	0.0417	0.5417	0.9921	419.255	0.9993
	rs9298463	52915555	0.9083	0.5722	0.0417	0.5417	0.9921	600.776	0.9998
	rs4259408	52921282	0.9083	0.5722	0.0417	0.5417	0.9918	303.701	0.9986
	rs11784921	52922009	0.9083	0.5722	0.0500	0.5327	0.9905	304.451	0.9987
	rs11777885	52922275	0.9083	0.5722	0.0417	0.5417	0.9918	297.022	0.9985
	rs6986984	52923371	0.9083	0.4444	0.0333	0.5439	0.9924	286.575	0.9984
	rs10096943	52923543	0.9083	0.5722	0.0417	0.5417	0.9918	274.069	0.9983
	rs4584139	52926708	0.9083	0.5722	0.0417	0.5417	0.9918	258.376	0.9981

Table XVI. Three regions and SNPs which had highest EHHT scores, strongest differentiations, and high derived allele frequency in the tested population (> 0.5) on chromosome 10 of the HapMap Phase II data. * marked new regions; **abbreviations**: Chrms — Chromosome, Popu. — Population, Pct — percentile. # marked region which was close to a candidate region found in Table 1, Sabeti et al. (2007). In the first column, the **Genes** provided names and positions of genes which were located in a region.

Chromosome, Tested Popu.	**SNP Name**	**Position**	**Derived Allele Frequency**			F_{st} **Score**	**Pct of** F_{st}	**EHHT Score**	**Pct of EHHT**
			CEU	**CHB+JPT**	**YRI**				
Chrms 10, CEU	rs10903907	3024509	0.9417	0.9222	0.1333	0.7120	0.9994	123.177	0.9951
	rs12782126	3026072	0.9333	0.9667	0.2333	0.6600	0.9985	123.316	0.9951
	rs11251627	3028967	0.9083	0.9333	0.0833	0.7515	0.9998	122.611	0.9950
	rs10794981	3031809	0.9083	0.9333	0.2583	0.5649	0.9929	110.887	0.9942
	rs7911099	3034806	0.9083	0.9444	0.2000	0.6421	0.9979	101.682	0.9937
	rs7914892	3035244	0.9083	0.9333	0.2000	0.6272	0.9973	98.094	0.9934
	rs10903915	3035615	0.9083	0.9500	0.2000	0.6497	0.9981	83.167	0.9915
	rs11251634	3036588	0.9083	0.9333	0.0583	0.7780	0.9999	80.906	0.9911
	rs10794985	3037529	0.9083	0.9333	0.2250	0.6005	0.9955	78.628	0.9907
	rs10903916	3037704	0.9000	0.9333	0.2000	0.6205	0.9969	76.557	0.9904
Chrms 10, CEU*,#	rs12262786	23929695	0.8167	0.9389	0.1917	0.5804	0.9942	349.148	0.9990
	rs7919210	23934980	0.8250	0.9389	0.2250	0.5496	0.9914	251.366	0.9983
	rs2050610	23935872	0.8167	0.9333	0.1917	0.5730	0.9936	205.902	0.9978
	rs6482309	23942942	0.8583	0.9389	0.1833	0.6152	0.9966	100.332	0.9936
	rs7100887	23943388	0.8417	0.9389	0.1833	0.6043	0.9958	112.922	0.9943
	rs1415423	23945789	0.8583	0.9389	0.1833	0.6152	0.9965	119.354	0.9947
	rs1415421	23945921	0.8583	0.9389	0.1917	0.6063	0.9960	131.722	0.9954
Chrms 10, CHB+JPT *PCDH15* (55,233-56,231kb)	rs10825242	55554795	0.2333	0.9167	0.1833	0.5955	0.9951	98.568	0.9956
	rs12218327	55561007	0.2250	0.9000	0.1250	0.6222	0.9970	93.558	0.9954
	rs9787578	55561682	0.3250	0.9444	0.2167	0.5606	0.9925	93.514	0.9954
	rs9787465	55562067	0.2250	0.9056	0.2167	0.5626	0.9927	91.449	0.9953
	rs4447073	55568150	0.3250	0.9444	0.2167	0.5606	0.9925	85.253	0.9950
	rs11004104	55588365	0.2333	0.9444	0.1333	0.6706	0.9987	78.291	0.9940
	rs11004105	55589349	0.2333	0.9444	0.1250	0.6771	0.9989	75.405	0.9935
	rs4636568	55590392	0.2333	0.9444	0.1333	0.6706	0.9987	73.598	0.9932
	rs11004106	55591311	0.2333	0.9444	0.1333	0.6706	0.9987	72.372	0.9929
	rs11004107	55591322	0.2250	0.9167	0.1083	0.6576	0.9984	67.084	0.9918
	rs10763079	55605400	0.6917	0.9889	0.2417	0.5507	0.9914	64.178	0.9911
	rs4935502	55625450	0.1583	0.8944	0.1667	0.6269	0.9973	121.242	0.9969
	rs10825275	55640296	0.9417	0.9944	0.3167	0.6264	0.9973	225.536	0.9996
	rs7093540	55641050	0.3417	0.9000	0.1333	0.5585	0.9922	224.442	0.9996
	rs11004141	55641433	0.1583	0.8667	0.1000	0.6411	0.9978	223.228	0.9996
	rs4935104	55646474	0.1500	0.8556	0.1833	0.5711	0.9934	209.818	0.9994
	rs1970519	55790948	0.2250	0.8667	0.0750	0.6210	0.9969	157.962	0.9983
	rs2028440	55796208	0.2333	0.8722	0.0750	0.6235	0.9970	137.462	0.9976
	rs11004267	55806859	0.2417	0.8667	0.0417	0.6424	0.9979	81.592	0.9944
	rs11004270	55817364	0.2250	0.8667	0.0417	0.6511	0.9982	78.882	0.9941
	rs11004275	55822846	0.2333	0.8667	0.0917	0.6018	0.9956	72.757	0.9930
	rs10825320	55826240	0.8583	0.9722	0.2583	0.5829	0.9944	72.754	0.9930
	rs2050998	55827170	0.8583	0.9722	0.2583	0.5829	0.9944	72.410	0.9929
	rs2050999	55827214	0.2333	0.8667	0.0917	0.6018	0.9956	70.715	0.9927
	rs9943342	55832701	0.2333	0.8667	0.0917	0.6018	0.9956	68.106	0.9921
	rs2795918	55847021	0.3250	0.9222	0.1167	0.6094	0.9962	65.802	0.9915
	rs1219862	55860311	0.3250	0.9222	0.0917	0.6313	0.9975	82.791	0.9945

Table XVII. Two regions and SNPs which had highest EHHT scores, strongest differentiations, and high derived allele frequency in the tested population (> 0.5) on chromosome 11 of the HapMap Phase II data. * marked new regions; **abbreviations**: Chrms — Chromosome, Popu. — Population, Pct — percentile.

Chromosome, Tested Popu.	**SNP Name**	**Position**	**Derived Allele Frequency**			F_{st} **Score**	**Pct of** F_{st}	**EHHT Score**	**Pct of EHHT**
			CEU	**CHB+JPT**	**YRI**				
Chrms 11, CHB+JPT*	rs821006	38066457	0.9750	0.9500	0.3417	0.5633	0.9944	74.432	0.9934
	rs821011	38082143	0.9833	0.9500	0.3583	0.5551	0.9935	71.546	0.9929
	rs820905	38137367	0.9833	0.9611	0.3167	0.6152	0.9972	147.722	0.9983
	rs898911	38154147	0.9833	0.9611	0.3083	0.6238	0.9980	98.601	0.9961
	rs10742415	38164195	0.9833	0.9833	0.3583	0.6092	0.9971	114.962	0.9974
	rs1381587	38165693	0.9833	0.9833	0.3667	0.6007	0.9968	114.432	0.9973
	rs6484977	38168478	0.9833	0.9833	0.3667	0.6007	0.9967	113.742	0.9973
	rs4755483	38176296	0.9833	0.9833	0.3667	0.6007	0.9967	112.905	0.9973
	rs4756422	38179004	0.9833	0.9833	0.3583	0.6092	0.9971	93.202	0.9955
	rs10836939	38180677	0.9833	0.9833	0.3583	0.6092	0.9971	93.267	0.9955
	rs10768362	38182171	0.9833	0.9833	0.3667	0.6007	0.9967	93.162	0.9955
	rs4756423	38182413	0.9833	0.9833	0.4167	0.5496	0.9929	92.888	0.9954
	rs10836940	38195060	0.9833	0.9833	0.3500	0.6176	0.9977	83.280	0.9947
	rs1462245	38242568	0.9917	0.9833	0.4333	0.5428	0.9923	535.322	0.9997
	rs2045737	38263161	0.9833	0.8389	0.1917	0.5978	0.9966	58.564	0.9902
Chrms 11, CEU*	rs10742415	38164195	0.9833	0.9833	0.3583	0.6092	0.9971	89.663	0.9933
	rs1381587	38165693	0.9833	0.9833	0.3667	0.6007	0.9968	97.501	0.9946
	rs6484977	38168478	0.9833	0.9833	0.3667	0.6007	0.9967	108.643	0.9958
	rs4755483	38176296	0.9833	0.9833	0.3667	0.6007	0.9967	119.824	0.9965
	rs4756422	38179004	0.9833	0.9833	0.3583	0.6092	0.9971	117.061	0.9964
	rs10836939	38180677	0.9833	0.9833	0.3583	0.6092	0.9971	122.364	0.9966
	rs10768362	38182171	0.9833	0.9833	0.3667	0.6007	0.9967	123.988	0.9967
	rs4756423	38182413	0.9833	0.9833	0.4167	0.5496	0.9929	124.806	0.9967
	rs10836940	38195060	0.9833	0.9833	0.3500	0.6176	0.9977	161.221	0.9977
	rs1462245	38242568	0.9917	0.9833	0.4333	0.5428	0.9923	90.075	0.9935
	rs1585555	38244083	0.9917	0.9833	0.4333	0.5428	0.9923	90.794	0.9937
	rs1599564	38244690	0.9917	0.9833	0.4333	0.5428	0.9922	90.319	0.9935
	rs2045737	38263161	0.9833	0.8389	0.1917	0.5978	0.9966	86.314	0.9927
	rs11034713	38391328	0.6667	0.0556	0.0333	0.5550	0.9935	469.665	0.9995
	rs12786969	38395092	0.6667	0.0444	0.0083	0.5989	0.9966	418.250	0.9994
	rs1435156	38448731	0.6833	0.0611	0.0083	0.5914	0.9963	86.225	0.9927

Table XVIII. One region and SNPs which had highest EHHT scores, strongest differentiations, and high derived allele frequency in the tested population (> 0.5) on chromosome 12 of the HapMap Phase II data. * marked new regions; **abbreviations**: Chrms — Chromosome, Popu. — Population, Pct — percentile. In the first column, the **Genes** provided names and positions of genes which were located in a region.

Chromosome, Tested Popu.	**SNP Name**	**Position**	**Derived Allele Frequency**			F_{st} **Score**	**Pct of** F_{st}	**EHHT Score**	**Pct of EHHT**
			CEU	**CHB+JPT**	**YRI**				
Chrms 12,	rs2407788	42674621	0.8750	0.8611	0.1833	0.5348	0.9907	44.893	0.9961
CHB+JPT*	rs842199	42743839	0.8167	0.8722	0.1583	0.5355	0.9908	38.057	0.9942
TMEM117	rs7971340	42762284	0.9167	0.9833	0.3000	0.6018	0.9961	44.539	0.9960
(42,516-43,070kb)	rs2638853	42763366	0.9167	0.8556	0.1333	0.6163	0.9971	41.135	0.9951
	rs17121310	42771080	0.9167	0.8556	0.2000	0.5437	0.9918	41.819	0.9953
	rs1352935	42771425	0.9083	0.9556	0.3000	0.5519	0.9924	45.644	0.9963
	rs7967957	42775258	0.9167	0.9556	0.2667	0.5941	0.9957	46.170	0.9964
	rs6582498	42782123	0.9167	0.9556	0.3000	0.5589	0.9930	47.269	0.9968
	rs17094092	42794527	0.9167	0.8556	0.2000	0.5437	0.9918	46.661	0.9966
	rs12315961	42796636	0.9167	0.9556	0.2750	0.5853	0.9953	50.398	0.9975
	rs11182423	42807802	0.9167	0.8556	0.1917	0.5527	0.9924	45.747	0.9963
	rs1643429	42879128	0.9167	0.8556	0.1250	0.6254	0.9975	42.564	0.9955

Table XIX. One region and SNPs which had highest EHHT scores, strongest differentiations, and high derived allele frequency in the tested population (> 0.5) on chromosome 13 of the HapMap Phase II data. * marked new regions; **abbreviations**: Chrms — Chromosome, Popu. — Population, Pct — percentile.

Chromosome, Tested Popu.	**SNP Name**	**Position**	**Derived Allele Frequency**			F_{st} **Score**	**Pct of** F_{st}	**EHHT Score**	**Pct of EHHT**
			CEU	**CHB+JPT**	**YRI**				
Chrms 13,	rs4310736	63174790	0.8250	0.9389	0.2583	0.5142	0.9907	41.525	0.9900
CHB+JPT*	rs9539851	63177581	0.8250	0.9389	0.2583	0.5142	0.9907	42.741	0.9906
	rs9539854	63181695	0.8333	0.9944	0.3417	0.5161	0.9909	45.613	0.9915
	rs9539855	63182064	0.8333	0.9944	0.3417	0.5161	0.9909	45.705	0.9915
	rs11843593	63185384	0.8333	0.9889	0.2833	0.5668	0.9946	45.658	0.9915
	rs9571012	63201277	0.9833	0.9444	0.2417	0.6679	0.9983	44.623	0.9912
	rs7985049	63203415	0.9917	0.9333	0.2833	0.6182	0.9973	45.249	0.9914
	rs4363749	63205262	0.9917	0.9333	0.2667	0.6354	0.9977	45.043	0.9913
	rs2807121	63286791	0.8250	0.9333	0.2417	0.5244	0.9916	47.596	0.9919
	rs536914	63320940	0.9833	0.9333	0.2667	0.6265	0.9975	53.095	0.9934
	rs621541	63321216	0.9833	0.9333	0.2667	0.6265	0.9974	53.169	0.9934
	rs1686806	63322252	0.9750	0.9333	0.2667	0.6178	0.9972	53.202	0.9934
	rs275913	63330923	0.9833	0.9333	0.2667	0.6265	0.9974	55.027	0.9942
	rs12428552	63336180	0.9833	0.9333	0.2667	0.6265	0.9975	54.179	0.9939
	rs824764	63338626	0.9833	0.9333	0.2750	0.6178	0.9973	53.395	0.9935
	rs1095072	63347162	0.9833	0.9389	0.2667	0.6342	0.9976	54.375	0.9940
	rs605874	63348614	0.9833	0.9389	0.2667	0.6342	0.9976	53.93	0.9938

Table XX. One region and SNPs which had highest EHHT scores, strongest differentiations, and high derived allele frequency in the tested population (> 0.5) on chromosome 15 of the HapMap Phase II data.

Chromosome, Tested Popu.	**SNP Name**	**Position**	**Derived Allele Frequency**			F_{st} **Score**	**Pct of** F_{st}	**EHHT Score**	**Pct of EHHT**
			CEU	**CHB+JPT**	**YRI**				
Chrms 15, CHB+JPT	rs11635593	61799912	0.1167	0.9167	0.1917	0.6708	0.9967	96.004	0.9995
	rs7163401	61811844	0.0583	0.9167	0.1917	0.7184	0.9984	90.311	0.9994
	rs7172848	61818807	0.0583	0.9167	0.1917	0.7184	0.9984	86.486	0.9993
	rs7164301	61821430	0.0583	0.9167	0.1917	0.7184	0.9984	84.642	0.9993
	rs8037083	61853580	0.0583	0.9167	0.2500	0.6868	0.9972	53.122	0.9982
	rs2414823	61860806	0.0583	0.9167	0.1917	0.7184	0.9984	49.036	0.9975
	rs7178111	61862564	0.1000	0.9167	0.1833	0.6891	0.9972	48.503	0.9974
	rs6494433	61869963	0.0583	0.9167	0.1917	0.7184	0.9984	47.406	0.9971
	rs4984317	61871710	0.0583	0.9167	0.1917	0.7184	0.9984	46.265	0.9968
	rs4984318	61871736	0.0583	0.9167	0.1917	0.7184	0.9984	45.711	0.9967
	rs8027701	61878102	0.0667	0.9167	0.1917	0.7114	0.9979	43.451	0.9959
	rs6494436	61878357	0.0583	0.9167	0.2000	0.7136	0.9980	42.897	0.9956
	rs4776681	61886816	0.0583	0.9167	0.1917	0.7184	0.9984	39.216	0.9939
	rs7173437	61887845	0.0583	0.8833	0.1917	0.6760	0.9968	38.332	0.9936
	rs2099921	61890902	0.0583	0.8833	0.1917	0.6760	0.9968	37.811	0.9934
	rs7178104	61892636	0.0583	0.8889	0.1917	0.6830	0.9971	35.796	0.9926
	rs7182375	61892683	0.0583	0.8889	0.3333	0.6153	0.9942	35.312	0.9924
	rs7165577	61904056	0.0583	0.8889	0.1833	0.6880	0.9972	33.160	0.9909
	rs2053593	61905916	0.0583	0.8889	0.1917	0.6830	0.9971	32.275	0.9901

Table XXI. One region and SNPs which had highest EHHT scores, strongest differentiations, and high derived allele frequency in the tested population (> 0.5) on chromosome 17 of the HapMap Phase II data. * marked new regions; **abbreviations**: Chrms — Chromosome, Popu. — Population, Pct — percentile. # marked region which was close to a candidate region found in Table 1, Sabeti et al. (2007). In the first column, the **Genes** provided names and positions of genes which were located in a region.

Chromosome, Tested Popu.	**SNP Name**	**Position**	**Derived Allele Frequency**			F_{st} **Score**	**Pct of** F_{st}	**EHHT Score**	**Pct of EHHT**
			CEU	**CHB+JPT**	**YRI**				
Chrms 17, CEU*,# *CA4* (55,582-55,592kb)	rs345187	55588298	0.9333	0.9389	0.1750	0.6821	0.9972	48.386	0.9956
	rs345184	55594073	0.9417	0.9278	0.1667	0.6837	0.9972	50.526	0.9961
	rs2452542	55600976	0.8500	0.7278	0.0333	0.5719	0.9910	49.052	0.9958
	rs9891296	55630776	0.9500	0.9222	0.2750	0.5694	0.9908	53.405	0.9967
	rs9893536	55663144	0.9500	0.9278	0.2667	0.5856	0.9921	51.670	0.9963
	rs8080640	55671729	0.9500	0.9278	0.2750	0.5767	0.9914	49.539	0.9959
	rs7216914	55678847	0.9500	0.9278	0.2000	0.6561	0.9963	48.490	0.9957
	rs237967	55686029	0.9500	0.9278	0.2750	0.5767	0.9914	44.898	0.9947
	rs237956	55695726	0.9500	0.9278	0.2667	0.5856	0.9921	43.142	0.9938
	rs237954	55698601	0.9500	0.9278	0.1917	0.6649	0.9965	40.970	0.9934

Table XXII. Maximum extended haplotype-based homozygosity test (EHHT) scores of the HapMap Phase II data of three population across human genome, chromosome 1 to chromosome 11.

Chromosome	**Sample**	**SNP**	**Position**	**Scores**
chromosome 1	CEU	rs9287131	192304513	1385.916
	CHB+JPT	rs10493514	73153100	5096.607
	YRI	rs2292275	161558841	72.873
chromosome 2	CEU	rs6739328	21650555	9178.164
	CEU	rs6547423	21651477	9178.164
	CEU	rs1477471	21651840	9178.164
	CHB+JPT	rs7594350	72425722	30805.837
	CHB+JPT	rs1400582	72474987	30805.837
	YRI	rs11883730	63445734	213.353
chromosome 3	CEU	rs341770	8954779	1345.997
	CHB+JPT	rs1386675	97865566	5371.473
	YRI	rs6763004	164851217	115.185
chromosome 4	CEU	rs1455724	60996026	1747.229
	CHB+JPT	rs6531808	86381856	2922.449
	YRI	rs4834160	128144330	53.851
chromosome 5	CEU	rs1054020	145505341	1557.327
	CHB+JPT	rs13182616	155717380	1649.290
	YRI	rs2431218	79905228	70.966
chromosome 6	CEU	rs3010521	48715542	628.272
	CHB+JPT	rs12202591	130680197	3719.419
	YRI	rs1264567	30474079	133.707
chromosome 7	CEU	rs17169262	136890892	8827.805
	CHB+JPT	rs11979882	90245236	1205.201
	YRI	rs569862	54641039	65.130
chromosome 8	CEU	rs2205153	91466385	1570.848
	CHB+JPT	rs13260166	50721620	1296.325
	YRI	rs16915414	51870380	87.408
chromosome 9	CEU	rs13291813	106392446	1408.915
	CHB+JPT	rs17810391	105983919	6384.794
	YRI	rs12004563	71920439	100.471
chromosome 10	CEU	rs7100458	74673099	5067.346
	CHB+JPT	rs2067732	86973092	280.070
	YRI	rs4623821	58998686	53.969
chromosome 11	CEU	rs11035187	39228629	3991.305
	CHB+JPT	rs2618296	37916220	8101.077
	YRI	rs4543965	37931669	64.632

Table XXIII. Maximum extended haplotype-based homozygosity test (EHHT) scores of the HapMap Phase II data of three population across human genome, chromosome 12 to chromosome 22.

Chromosome	**Sample**	**SNP**	**Position**	**Scores**
chromosome 12	CEU	rs4149160	20905844	950.954
	CHB+JPT	rs10863071	84669944	4077.948
	YRI	rs12312066	66002935	94.098
chromosome 13	CEU	rs2231332	37822263	3292.477
	CHB+JPT	rs9506383	19282447	7783.522
	YRI	rs4884166	54667146	124.023
chromosome 14	CEU	rs1253642	51496904	1138.814
	CHB+JPT	rs10136790	67471698	5801.927
	CHB+JPT	rs10133262	67477691	5801.927
	YRI	rs7156228	59748861	58.875
chromosome 15	CEU	rs4404024	69734324	1101.560
	CHB+JPT	rs1472946	49011221	695.312
	YRI	rs12939	40363868	62.413
chromosome 16	CEU	rs2245201	77092871	235.138
	CHB+JPT	rs16969790	20173234	200.338
	YRI	rs2157854	22851261	96.901
chromosome 17	CEU	rs12450486	41620562	489.866
	CHB+JPT	rs12051550	20070138	832.638
	YRI	rs17175543	54158597	33.747
chromosome 18	CEU	rs9319771	64871413	4272.775
	CEU	rs10432227	64871622	4272.775
	CHB+JPT	rs2222394	18502286	372.659
	YRI	rs11664364	36269816	76.499
chromosome 19	CEU	rs978348	34008970	335.706
	CHB+JPT	rs16968285	33056093	224.119
	YRI	rs10422765	23730242	54.129
chromosome 20	CEU	rs184147	37751486	452.847
	CHB+JPT	rs4611702	40670251	293.510
	YRI	rs290429	52108753	63.020
chromosome 21	CEU	rs2027717	23329336	85.046
	CHB+JPT	rs2834997	35860466	826.522
	YRI	rs2070865	39637389	33.553
chromosome 22	CEU	rs5999761	33929534	1034.820
	CHB+JPT	rs17377643	40482934	260.948
	YRI	rs8142666	28754601	56.552

CHAPTER IV

SUMMARY AND CONCLUSIONS

We propose three new score tests, EGHT, HMMT and EHHT to detect recent selection via examining the extent of haplotype homozygosity in genome-wide scans. They share a common test statistic but postulate different null hypotheses. Intuitively, EGHT may show significant results if either HWE or LE is invalid, HMMT could do so if either HWE is invalid or there exists higher order LD interaction than pair-wise ones among SNPs, and EHHT provides high scores only when haplotype version of HWE is invalid in a chromosome region. Roughly speaking, EGHT and EHHT are two extremes: EGHT tends to reject the null too often and detect too many signals given high density of SNP data and so the presence of LD is a fact of ubiquity, EHHT is the most conservative one since it only gives high scores in the presence of excess homozygosity. We start from a measure T of extent of homozygosity, and then provide the distribution of T and its mean and variance under the null hypothesis of each test case (**METHODS**). This facilitates the calculation of our test statistics.

By simulating data under the null hypothesis of the EGHT, we evaluate the robustness of the three tests in terms of false positive rates and confirm that the EHHT is the most robust. We generate samples with coalescent programs, such as SelSim and ms, to study the performance of the EHHT. It's worthy mention that the existing popular tests usually do not follow a clear distribution. The EHHT, however, is asymptotically normal which makes analysis and applications easier. We apply these tests to HapMap Phase II data for genome-wide screen, compare with previously reported candidate regions, and search for new candidate regions based on

high EHHT scores and population differentiations. It is encouraging that our EHHT scores confirm 17 regions of excess homozygosity out of 20 candidates reported by Sabeti et al. (2007). The statistics also validate the relative demographic history of the African, European, and east Asian populations. Our plots suggest multiple regions of excess homozygosity.

In summary, the main contributions are: we show that the EHHT could be used to detect regions of excess homozygosity, which could be candidates of recent selection for further investigations by additional requirement such as the criteria used in Sabeti et al. (2007): selected alleles are newly arisen, likely to be highly differentiated among populations, and they also have biological effects. The EHHT is conservative and robust. Comparing with the existing popular methods, the EHHT performs just as well or even better. Moreover, the EHHT is straightforward and asymptotically normal. In addition to the EHHT, we show that the other two methods, EGHT and HMMT, are useful in genome-wide scans for a general pictures of the strength of LD and violation of HWE by comparing test scores of different population samples. For candidate regions which have selection signals, the comparison of the three test scores might provide clues of either LD or violation of HWE or both which lead to high test scores.

Due to the conservative nature of the EHHT, one might miss some candidate regions in which HWE is roughly valid but there exists LD. Thus, high EHHT scores are not a sufficient and necessary condition for detection of selection signal. Notwithstanding, the EHHT could be a new tool in addition to existing methods of detecting selection. Population geneticists have proposed several tests for inferring a selective sweep. Jensen et al. (2005) summarized the most important tests, including ones based on increased LD (Przeworski, 2002; Kim and Nielsen, 2004). We like the current statistics because they exploit dense SNP genotyping and depend on minimal

assumptions. Of course, the lack of a detailed model has its disadvantages. For example, our tests say nothing about the age of a favorable mutation. This issue is obviously intertwined with variations in recombination rates across the genome.

REFERENCES

Akey JM, Zhang G, Zhang K, Jin L, Shriver MD (2002) Interrogating a high-density SNP map for signature of natural selection. Genome Res 12: 1805-1814.

Akey JM, Eberle MA, Rieder MJ, Carlson CS, Shriver MD, Nickerson DA, Kruglyak L (2004) Population history and natural selection shape patterns of genetic variation in 132 genes. PLoS Biol 2: e286.

Ayers KL, Lange K (2008) Penalized estimation of haplotype frequencies. Bioinformatics 24: 1596-1602.

Bamshad M, Wooding SP (2003) Signatures of natural selection in the human genome. Nat Rev Genet 4: 99-111.

Bersaglieri T, Sabeti PC, Patterson N, Vanderploeg T, Schaffner SF, Drake JA, Rhodes M, Reich DE, Hirschhorn JN (2004) Genetic signatures of strong recent positive selection at the lactase gene. Am J Hum Genet 74: 1111-1120.

Enattah NS, Sahi T, Savilahti E, Terwilliger JD, Peltonen L, Järvelä I (2002) Identification of a variant associated with adult-type hypolactasia. Nature Genet 30: 233-237.

Fay JC, Wu CI (2000) Hitchhiking under positive Darwinian selection. Genetics 155: 1405-1413.

Fu YX, Li WH (1993) Statistical tests of neutrality of mutations. Genetics 133: 693-709.

Hanchard NA, Rockett KA, Spencer C, Coop G, Pinder M, Jallow M, Kimber M, McVean G, Mott R, Kwiatkowski DP (2006) Screening for recently selected alleles by analysis of human haplotype similarity. Am J Hum Genet 78: 153-159.

Hudson RR, Bailey K, Skarecky D, Kwiatowski J, Ayala FJ (1994) Evidence for Positive Selection in the Superoxide Dismutase (Sod) Region of Drosophila melanogaster. Genetics 136: 1329-1340.

Hudson RR (2002) Generating samples under a Wright-Fisher neutral model. Bioinformatics 18: 337-8.

Kim Y, Nielsen R (2004) Linkage disequilibrium as a signature of selective sweeps. Genetics 167: 1513-1524.

Lamason RL, Mohideen MA, Mest JR, Wong AC, Norton HL, Aros MC, Jurynec MJ, Mao X, Humphreville VR, Humbert JE, Sinha S, Moore JL, Jagadeeswaran P, Zhao W, Ning G, Makalowska I, McKeigue PM, O'donnell D, Kittles R, Parra EJ, Mangini NJ, Grunwald DJ, Shriver MD, Canfield VA, Cheng KC (2005) SLC24A5, a putative cation exchanger, affects pigmentation in Zebrafish and humans. Science 310: 1782-1786.

Long JC, Williams RC, Urbanek M (1995) An E-M algorithm and testing strategy for multiple-locus haplotypes. Am J Hum Genet 56: 799-810.

Poulter M (2003) The causal element for the lactase persistence/non-persistence polymorphism is located in a 1 Mb region of linkage disequilibrium in Europeans. Ann Hum Genet 67: 298-311.

Przeworski M (2002) The signature of positive selection at randomly chosen loci. Genetics 160: 1179-1189.

Ronald J, Akey JM (2005) Genome-wide scans for loci under selection in humans. Hum Genomics 2: 113-125.

Sabeti PC, Reich DE, Higgins JM, Levine HZ, Richter DJ, *et al.* (2002) Detecting recent positive selection in the human genome from haplotype structure. Nature 419: 832-837.

Sabeti PC, Varilly P, Fry B, Lohmueller J, Hostetter E, *et al.* (2007) Genome-wide detection and characterization of positive selection in human populations. Nature 449: 913-918.

Spencer CCA, Coop G (2004) SelSim: a program to simulate population genetic data with natural selection and recombination. Bioinformatics 20: 3673-3675.

Tajima F (1989) Statistical methods for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123: 585-595.

The International HapMap Consortium. (2007) A second generation human haplotype map of over 3.1 million SNPs. Nature 449: 851-861.

Vallender EJ, Lahn BT (2004) Positive selection on the human genome. Hum Mol Genet 13 (Spec. No. 2): R245-R254.

Voight BF, Kudaravalli S, Wen X, Pritchard JK (2006) A map of recent positive selection in the human genome. PLoS Biol 4: e72.

Weir BS, Cockerham CC (1984) Estimating F-statistics for the analysis of population structure. Evolution 38: 1358-1370.

VITA

Ming Zhong was born in Anhui, a south province of China. In 1995, He graduated from Lujiang High School and entered the Special Class for the Gifted Young (SCGY) program in University of Science & Technology of China. He received his bachelor degree in Physics in July 2000. He was awarded Master degrees in Physics and Statistics from Texas A&M University, in August 2005 and May 2007, respectively. He continued his studies under the supervision of Dr. Ruzong Fan and was bestowed a Doctor of Philosophy in Statistics from Texas A&M University in May 2010. The title of his dissertation is *Extended Homozygosity Score Tests to Detect Positive Selection in Genome-wide Scans.* His correspondence address is Department of Statistics, Texas A&M University, 3143 TAMU, College Station, TX 77843-3143 or by email: zhongming@stat.tamu.edu